Towards Entity Status

Inauguraldissertation
zur
Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität
zu Bonn

vorgelegt von
Maria Klara Wolters
aus
Rheydt-Giesenkirchen

Bonn 2001

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Berichterstatter:	Prof. Dr. Winfried Lenders
2. Berichterstatter:	Prof. Dr. Wolfgang Hess
Tag der mündlichen Prüfung:	8. November 2000

Acknowledgements

When I began this work in December 1997, I planned to develop a corpus-based prosody module for concept-to-speech synthesis. I started with the prosodic marking of given and new information. In September 2000, I completed a thesis which is largely inspired by the headaches that givenness has caused me.

First of all, I would like to thank Wolfgang Hess for giving me the freedom to pursue this research, and for his constant encouragement. That I was able to finish in three years is mostly due to three people. Carlos Gussenhoven helped me turn a house of cards into a thesis. Michael Strube collaborated with me on the research presented in Chapter 7. Last, but not least, in the final months, Winfried Lenders made sure I did not stray too far from the path of my research.

During various extended lunches, Bernhard Schröder and I talked about semantics, semiotics, and linguistics. Edda Leopold discussed previous stages of Section 5.4 with me. Johann Juchem patiently introduced me to the work of Gerold Ungeheuer.

My work on the Boston Radio News Corpus would not have been possible without the cooperation and hospitality of the Induction of Linguistic Knowledge group at Tilburg. In particular, I would like to thank Sabine Buchholz (+ Wouter), Bertjan Busser, Antal van den Bosch, and Walter Daelemans. Jörg Meyer, IMS Stuttgart, kindly provided the Stuttgart Radio News Corpus.

A big thank you to Simone Teufel for putting up with me during a working holiday at the University of Edinburgh in March 1999; I hope it's been worth the trouble. In July 1999, I had the privilege of three weeks at Stanford, working with David Beaver, Brady Zack Clark and Edward Flemming, and spending quality time with Juno Nakamura and Angie Kortenhoven. Thank you, David, for making this possible. In the week before that, Michael Strube organized a one-week stay at the Institute for Research in Cognitive Science, University of Pennsylvania. I would like to thank IRCS for its hospitality.

The audiences of the ILK Colloquium at Tilburg, the Phonetics-and-Phonology-Workshop at Edinburgh, the Linguistics Colloquium at Stanford University, the Arbeitskreis Linguistische Pragmatik at the 25th Annual Meeting of the German Society of Linguistics (Deutsche Gesellschaft für Sprachwissenschaft), Konstanz, 1999, and the Prosodiegroep at Nijmegen provided valuable comments on earlier stages of Chapter 6 during a time when I still thought I was going to write a thesis about prosody.

For their comments on ideas for or aspects of present and past instantiations of this thesis, I would like to thank, apart from those I have already mentioned, Ralf Backhausen, Ellen Bard, Janet Cahn, Paul Dekker, John Fry, Stefan Kaufmann, Ewan Klein, Emiel Krahmer, Jonas Kuhn, Bob Ladd, Kathy McCoy, Margo Melnicove, Massimo Poesio, Thomas Portele, Toni Rietveld, Udo Stiehl, Marc Swerts, Bonnie Webber, and Henk Zeevat. Amit Almor, Kai Alter, David Beaver, Pia Bergmann, Thorsten Brants, Gerald Echterhoff, John Fry, Maria Luisa

García Lecumberri, Klaus von Heusinger, Joachim Jacobs, Brett Kessler, Jonas Kuhn, Stefan Krauss, Ivana Kruijff-Korbayová, Truus Kruyt, Massimo Poesio, Hector Ortiz Lira, Paul Piwek, Andreas Späth, and Thomas Ede Zimmermann kindly sent me papers and theses that I became interested in along the road. Angela Niederbäumer and Martin Volk provided me with NP-chunked data that did not find its way into this thesis for lack of time. It was great fun to work with David Beaver, Donna Byron, Brady Clark, Edward Flemming, Mathias Kirsten, Hansjörg Mixdorff, Bernhard Schröder, and Petra Wagner, although our joint papers or paper drafts never made it into the final thesis except for the odd citation. (Michael Strube was also fun to work with, but hey, he's already been acknowledged twice.)

Bernhard Schröder, Winfried Lenders, Petra Wagner, Dietmar Böhmer, Mathias Kirsten, Antal van den Bosch, Christina Widera and Florian Höfer read, proofread, and commented on drafts of various chapters and appendices—Petra, Christina, Antal, and Mathias on very short notice. Thomas Bender proofread near-final drafts of all chapters and of the German *Zusammenfassung*; Petra Wagner checked a draft of the bibliography. Wolfgang Hess, Winfried Lenders, David Beaver, and Miles Osborne commented on the submitted version. The thesis was revised at Rhetorical Systems, Edinburgh, and at the Department of Linguistics, Edinburgh University.

Thanks to my colleagues and ex-colleagues at the Institut für Kommunikationsforschung und Phonetik, Christina Widera, Michael Mennen, Gisela von Neffe, Dietmar Lancé, Gerd Willée, Petra Wagner, Anja Elsner, Julia Abresch, Stefan Breuer, Lioba Faust, Thomas Portele, Hans-Christian Schmitz, Gerit Sonntag, Karlheinz Stöber, and Volker Strom, as well as to my patient students, who had to wait for transcripts and supervision while I was busy finishing the thesis. Harald Ketzer and Thorsten Henrici were able and reliable student teaching assistants; Amanda Wildner helped with the measurements for a prosody experiment that did not survive the Second Great Redesign of May 2000.

Finally, I would not have survived this experience without the love of my fiancé, Thomas Bender, and of my parents, Roswitha and Heinz-Dieter Wolters. This thesis is dedicated to the memory of Julius the cat, an inquisitive little fellow, who has now ceased to be and is pushing up the daisies.

Contents

List of Tables

List of Figures

Conventions

Mathematical . . .

Probability and Statistics	
X	random variable
x	value of random variable
Xxxx	factor
p(X)	probability of X
p(X|Y)	probability of X given Y
L(C)	likelihood
General Linear Models	
X ∼	X is modelled by the formula on the right-hand side
Y + Z	the effects of Y and Z are added separately to the model
Y:Z	there is an interaction between Y and Z
Y*Z	Y + Z + Y:Z

. . . And Non-Mathematical

Citations: For all citations in German, I give English translations in *italics* in a footnote. Unless otherwise stated, these translations are my own.

Examples: In all examples, relevant expressions are enclosed in square brackets. Letters co-index co-specifying referring expressions.

A note on personal pronouns: I will usually refer to the communicator as “she” (mnemonic: She=Speaker), and to the addressee as “he” (mnemonic: He=Hearer).

Prologue

> Ich spreche auf diesen Seiten zwar über vieles, aber über noch mehr sage ich nichts, ganz einfach deshalb, weil ich darüber keine klaren Vorstellungen habe. Ein gutes Motto für mein Buch wäre darum ein Zitat von Boscoe Pertwee, einem (mir unbekannten) Autor des 18. Jahrhunderts, das ich bei Gregory (1981: 558) gefunden habe: "Früher war ich unentschieden, aber heute bin ich mir nicht mehr so sicher".[1] (Eco 2000, p. 9f.)

[1] *On these pages, I talk about a lot of things, but I do not say anything about even more things, simply because I do not have any clear ideas about them. Therefore, a good motto for my book would be a citation from Boscoe Pertwee, an eighteenth century author (unknown to me), which I found in Gregory (1981: 558): "Before, I was undecided, but today, I'm not so sure anymore".*

1 Introduction

Givenness is a term that haunts the linguistic literature. It is particularly persistent when researchers talk about referring expressions: Does that expression refer to something new? Or did the addressee know the referent already? Is that which is known topical, as well? Are there dimensions of oldness, scales of givenness, hierarchies of accessibility? New questions keep popping up like the heads of Hydra.

I do not propose to answer these questions here. This means that I will neither propose a theory of givenness, nor develop a scheme for annotating givenness in arbitrary texts. Rather, I propose to take a step back and inspect one of the Hydra's heads more closely: the givenness of discourse entities. More informally, discourse entities are what can be referred back to in discourse by a noun phrase; more formally, they are conceptual constructs that (computational) linguists use to model the semantics of referring expressions. The inspection proceeds in two steps:

Step 1: describe the factors that are involved in determining the givenness of discourse entities. This is the subject of Chapters 2. To avoid confusion with related ill-defined concepts such as "theme" or "coherence", I will summarise the web of factors that influence the givenness of a discourse entity under the heading *entity status*. Chapter 3 discusses the influence of thematicity, topicality, discourse structure, and coherence on entity status, while Chapter 4 relates entity status to salience, accessibility, and activation.

Step 2: examine how aspects of entity status can be measured in corpora, and how those aspects influence the form of referring expressions. This is the focus of the remainder of the thesis. Chapter 5 focuses on methodology, while Chapters 6–8 report on empirical studies.

This introduction is structured as follows: First, I explain why entity status was developed (Section 1.1), followed by general remarks on methodology in Section 1.2. Next, I outline the contributions of this thesis to the field of (computational) linguistics (Section 1.3) and finally, I give an overview of the thesis (Section 1.4).

1.1 Why Entity Status?

I started on the random walk that eventually became this thesis with the aim to build a corpus-based module for prosody generation in Content-to-Speech synthesis. I began with looking at the prosodic correlates of givenness. When I set out to search the literature for a theory of givenness on which I could base some corpus annotations, I found a swamp. Givenness appears to be a metaphor that researchers stretch and adapt as they like, depending on whether they

are looking for the cognitive substance of their linguistic intuitions, whether they want nice heuristics for their experiment design, or whether they want to explain something by the fact that something is known somehow to somebody. This proliferation of uses would fascinate sociologists of science. But it makes givenness somewhat less useful than it could be.

This experience led me to jettison givenness and hunt for a concept that was:

1. theory-independent enough to be compatible with different views on how discourse is structured, different models of speaker/hearer interaction, and different views on the cognitive processing of language,

2. precise enough so that it can be explicated easily in the framework of an adequate theory,

3. useful enough for describing and analysing discourse, even if it has not been embedded into a theory.

That concept is what I call *entity status*. Entity status is a structured bundle of information that collects all those properties of discourse entities which can conspire to make one entity more or less given than another. Entity status is an analytic construction; I would not claim that it as such is psychologically real, although some of the properties that it incorporates have been motivated by psycholinguistic research. The only strong claim that I make with entity status is the following: When talking about a complex gradient notion such as givenness, it does not make sense to let hierarchies and taxonomies of givenness square off against each other which cover different aspects of the same phenomenon. Instead, we should accept that we are dealing with a multi-faceted phenomenon and develop appropriate tools for its analysis.

The properties of a discourse entity that describe its status fall into two large classes:

Structural properties: This group describes the position of the discourse entity in the various levels of discourse structure, its connections to the various discourse segment topics, and its relation to other discourse entities.

Management properties: This group describes how the initial description of an entity is built when the discourse entity is first mentioned in a discourse. It monitors the entity's activation. Finally, it stores the links by which the entity can be accessed.

How the analytic construction of entity status is filled depends on the theoretical basis we choose. I will discuss several alternative instantiations of entity status in Chapters 3 and 4 passim. In Appendix A, entity status is illustrated by two sample analyses.

Some of the properties that affect the status of a discourse entity can be observed in corpora and examined with statistical methods. In the three empirical chapters, Chapter 6–7, I examine what these properties are, and how they can be analysed with traditional statistical methods from the fields of stochastic processes and generalised linear models. First, a case study of entity status in radio news (Chapter 6) reveals that if we want to observe some of the more interesting aspects of entity status, we first need a detailed analysis of the communication situation in which a text was created. Second, Chapter 7 explores the influence of that quantitative measure on pronominalisation.

1.2 Methodology

Another problem that I encountered in my work on prosodic correlates of givenness led me to question the role of statistics in my research. In the particular genre I looked at, radio news, prosodic correlates of givenness are not exactly straightforward. Since results are only good results if we have $p < 0.05$, and since it is impossible to argue statistically that something is not the case, I encountered innumerable "yes, but . . .", which do not spring to mind if your p-values are well-behaved. Clearly, statistics can help find patterns in data (exploratory statistics), summarise data quantitatively (descriptive statistics, statistical modelling), and determine whether associations found in one data set can be generalised to the population that the data comes from (inferential statistics). But the point is that it can only help you do the work of interpretation and analysis, it cannot do that work for you (Gigerenzer, Swijtink, Porter, Daston, Beatty and Krüger 1989). If your hypotheses are ill-defined, if your data does not fulfil certain preconditions, if you are trying to predict something which humans cannot even measure reliably, both you and your statistics are in deep trouble.

Another quandary lies in the demands that statistics makes on its users. In order to keep the mathematics manageable, many statistical methods make simplifying assumptions about the data that only a few real-world applications are likely to meet. Those methods that make only a few assumptions, such as most non-parametric analysis methods, are difficult to interpret. The way in which I will use statistical methods in this thesis is best described by the following quote:

> Statistical models are only approximations to the underlying data generating mechanism, but are, nevertheless, useful representations to aid us in understanding what is going on. In many cases, they may not even be at all realistic, but, when shown not to fit well to the data, indicate clearly that a certain mechanism is *not* operating. (Lindsay 1995, p. 97)

I do not claim to introduce new methods to corpus linguistics. Rather, I use tried and tested methods in an unfamiliar way—as a means for testing linguistic hypotheses and with clear caveats when the linguistic data was obviously not as well-behaved as the method would have liked it to be. Most of the methods I use are parametric: they make clear, strong assumptions about the distribution of the random variables that are to be modelled. The disadvantage is that some of these assumptions are bound to be too strong. But in contrast to non-parametric models, which make fewer assumptions, the results are much easier to interpret, as long as one bears in mind what simplifications had to be made in order to apply the model.

Statistical analysis of corpora has its limits. All it can do is to give us quantitative summaries of how an annotator reacted to the text she worked on—to the extent that her annotations reflect her reactions. For the sake of replicability, the annotations should be consistent across annotators: any well-trained annotator should react to the text with the same labels. This requirement is called *reliability* and has become increasingly important in corpus linguistics. But there are limits to what we can annotate reliably, and we hit these limits when we want people to label how a given string of words might be processed cognitively. This is a fundamental problem in any corpus study of referring expressions. I solve this problem in two ways. For the large-scale studies on the BROWN-COSPEC corpus, which is described in Appendix C, we identified sequences of referring expressions that access the same discourse entity and added little other complex information. Second, for the small-scale statistical studies reported in Chapter 6, I

work with the labels of one annotator only. These labels express her judgements about cognitive accessibility and the source of new discourse entities. The statistical analyses of this data only indirectly say something about the use of referring expressions; they merely summarise how the annotator reacted to the text. In order to determine to what extent properties of the text influence these reactions, the experiment would need to be replicated with different annotators.

The emphasis on cognitive models in the more functionally oriented literature not only leads to difficulties with corpus studies, as stated above. It also obscures the fact that language is a social semiotic (Firth 1950, Halliday 1978). I do not know of any linguistic theory that would integrate both aspects in an appealing framework, that can afford to present the complete, albeit complex picture, and I mistrust any attempt to reduce part of that complexity to one or two keywords plus maybe some affiliated maxims, elegant though they may appear.

Instead, I choose my analytical tools depending on the research goals: In the study of radio news, I adopt a functionalist perspective, which integrates results from media studies and cognitive science. When documenting and modelling patterns of language use, I use statistics.

1.3 Contributions

> Erschöpfende Belesenheit masse ich mir nicht an, und sie ist bei dem Umfange unserer Literatur kaum zu verlangen. Manches, was ich für mein Eigenstes halte, mag sich schon längst in den Werken Anderer vorfinden; und wenn ich es wirklich zum ersten Male zu Papier gebracht habe, so kann, ohne dass ich es mich entsinne, mein verewigter Vater der geistige Urheber gewesen sein.[1] (von der Gabelentz 1891, p. VI)

What the contributions of this study are depends to a large extent on the interests of the reader.

For those who are interested in theoretical debates, I offer entity status, a flexible concept that I have used as an extralinguistic basis for studying the form of referring expressions. It is relatively theory-neutral; I have explored its relation to several theories of discourse structure, such as Rhetorical Structure Theory (Mann and Thompson 1988), the discourse theory of Grosz and Sidner (1986) and van Dijk and Kintsch's (1983) propositional theory of text processing. Readers with a leaning towards semiotics might find the introduction to and application of Ungeheuer's (1987c) perspective on communication interesting, a perspective which has been ignored in most of computational linguistics.

For those who are interested in practical results, I offer three empirical studies: In the first study I analyse entity status in radio news. Radio news, and agency news in general, are a very popular genre in computational linguistics and speech processing: Audio mining and text classification algorithms are trained on such corpora, and large radio news corpora such as the Boston University Radio News Corpus (WBUR, Ostendorf, Price and Shattuck-Hufnagel 1995) and the Stuttgart Radio News Corpus (SRN, Rapp 1998) are used in the development of speech

[1] *I do not claim exhaustive scholarship, and such a scholarship can hardly be demanded given the size of our literature. Some of what I might deem my very own might possibly already be found in the works of others, and if I should really have written something down for the first time, my own deceased father may have been the originator, even though I cannot remember it.*

synthesis systems, in particular in developing prosody modules. What is new about this thesis, at least for that part of the linguistics community I come from, is that I attempt to take the genre seriously, that I draw on results from media studies to describe how the particular communication situation in radio news will affect the status of discourse entities, and that I show, using these results, how deeply problematic classifications such as "familiar" or "identifiable" can be—especially when the analyst leaves the cozy speaker/hearer dyad and moves on to more complex settings.

Second, I attempt to develop a statistical model of structural entity status based on sequences of referring expressions that co-specify the same discourse entity. These sequences consist of mentions of the same discourse entity. In each sequence, two states can be discerned, active and backgrounded. In order to model these patterns, I propose a model that consists of two interlinked stochastic processes, a non-stationary Poisson process, which generates the mentions of a discourse entity, and a Markov Chain, which simulates the switch between active and backgrounded state. As far as I know, this is the first such model. A preliminary evaluation indicates that the model falls short of the initial expectations for two reasons: Firstly, it fails to cover syntactic constraints on the occurrence of discourse entities, secondly, it does not explain how the co-text influences the probability that an entity will get mentioned. In order to extend it in these directions, we need both more sophisticated mathematical models and a corpus that is large enough to allow to estimate the influences of the co-text.

Finally, I study the connection between entity status and pronominalisation in a sizeable corpus of written British English, where entity status is operationalised via distance to last mention. I compare distance to six other factors, syntactic function, syntactic function of the antecedent, form of the antecedent, number of competing antecedents, sortal class, and agreement. As far as I know, this is one of the first systematic cross-genre studies of the factors that influence pronominalisation. I also develop a simple method on the basis of generalised linear models that helps detect powerful and robust features. Powerful features predict pronominalisation well, while the influence of robust features does not vary greatly with genre.

1.4 Overview

Before you leave the introduction, here is a road map of what is to follow. For easy reference, Figure 1.1 indicates dependencies between chapters by arrows.

Chapters 2, 3 and 4 form the theoretical part. In Chapter 2 I explain what entity status is and discuss its semiotic and communication theoretic foundations. The communication theory of Gerold Ungeheuer on which large sections of Chapter 2 are based is introduced in Appendix E. In Chapters 3 and 4 I review the literature on structural (3) and management (4) aspects of entity status, and link the concept to various theories—theories of discourse structure and thematicity in Chapter 3, and theories of activation and accessibility in Chapter 4.

Together with Appendix B, Chapter 5 forms a methodological interlude. In Chapter 5 I discuss schemes for annotating entity status and develop and motivate the annotation strategies that will be used in the empirical Chapters 6 and 7. In that chapter, I also explore distance from last mention as a measure of entity status and take the first steps towards a statistical model of co-specification sequences. The methods for statistical corpus analysis that I will use in these Chapters are described in more detail in Appendix B.

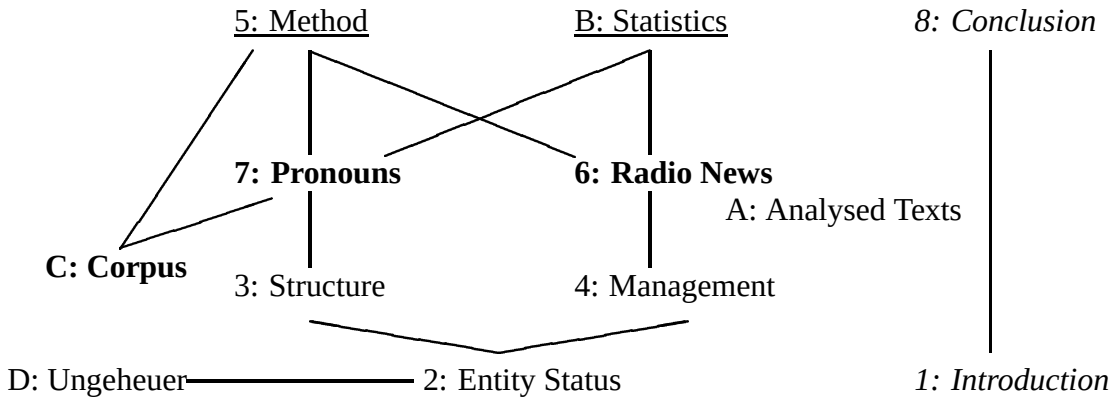


Figure 1.1. Overview of the thesis structure. Arrows indicate dependencies. X $\rightarrow$ Y means: X provides background information for Y. Numbers denote chapters, capital letters denote appendices. **Bold:** empirical part, underlined: methodological part, *italics:* introduction and conclusion

The empirical part consists of Chapters 6 and 7. In Chapter 6 I examine entity status in a small corpus of radio news and present several detailed text analyses. On the basis of a much larger corpus, I investigate influences on pronominalisation in Chapter 7.

Chapter 8 summarises the results of the thesis and outlines potential applications to prosody research. The Appendices contain two sample texts (Appendix A), a brief review of the necessary statistics, in particular generalised linear models (Appendix B), documentation for the BROWN-COSPEC-corpus together with the annotation manual for Sortal Classes (Appendix C), and a brief introduction to the communication semantics of Gerold Ungeheuer (Appendix D).

All chapters except for the introduction and the conclusion carry a summary of the main results, either as an explicit section or as a subsection. Readers who are only interested in the empirical results should read the summary of Chapter 2 and then move on to the chapter of their choice. More thorough readers might want to add the summaries of Chapters 3 and 4 to that trajectory. Readers who are looking for literature reviews or theoretical discussions should read Chapters 3 and 4; those who have a penchant for philosophical discussions should definitively try the complete Chapter 2 and Appendix D. Appendix E contains a summary in German.

2 What is Entity Status?

In this chapter, I will provide a theoretical foundation for the concept of entity status. When we link givenness or entity status to a theory of cognitive processing, in particular to an implemented model, we make our theory more specific, and hence more testable. This is the path that most scholars working on givenness strive to take. But the focus on cognitive models has its downsides: Firstly, quite a few researchers rely on folk-psychological theories of memory and cognition.Secondly, the more deeply enmeshed you become in a particular theory of language processing, the more you lose sight of communication. In particular, you tend to forget that, as Firth (1950) pointed out, language is used for communication in societies. In order to remember this fact, you need to return to a more fundamental way of thinking about language: communication theory. This is the path I will take in this chapter.

However, no matter what theory I will align myself with in the following pages, some reader will have good reason to criticise it on philosophical grounds. There is no generally accepted theory of human communication, nor is there "a" generally recognised semiotic theory. For the theory of communication, I largely rely on the work of Gerold Ungeheuer (1987c). Ungeheuer's perspective on communication is attractive because he takes pains not to simplify anything; this is probably the reason why he never developed a full-blown theory. Since his work is little known, even within Germany, Appendix D gives a brief introduction to his ideas.

This chapter is structured as follows: First, in Section 2.1, I discuss what discourse entities are, both from a semantic and from a semiotic point of view, and develop the concept of entity status in more detail. Next, in Section 2.2, I further develop the concepts of discourse entity and entity status on the basis of Gerold Ungeheuer's approach to communication. Section 2.3 concludes with a summary of the main points.

2.1 What is a Discourse Entity?

The notion of discourse entities is relatively young, but its source is very old: the problem of reference. Since the time of the Ancient Greeks, philosophers have debated if and how words can be used to refer to objects in the real world (Coseriu 1975). The technical term "discourse entity" was introduced in order to dissociate two research questions, namely "What is the nature of reference?" and "What does a referring expression mean?". To give an idea of the problems involved in thinking about reference, I will sketch the work of Frege and of Russell in Section 2.1.1). Then, in Section 2.1.2, I will contrast their results with the work of Karttunen, who invented the term "discourse referent" (Section 2.1.2), which is roughly the same as our "discourse entity". That term will be developed in more detail in Section 2.1.2.

2.1.1 Reference

Frege (1892)[1] was the first to distinguish between sense and reference. When we know the reference of an expression, we can evaluate whether it is true or false relative to those entities and concepts in the world that it refers to. Its sense, on the other hand, is the mode in which it is presented, how we think about and understand that expression. Although Frege establishes the notion of sense on the level of thoughts, a sense is not something that exists in the mind of an individual. Instead, the senses of signs are fixed and maintained by a community. Take for example the sentences in the following example, cited after (Frege 1892, page 28):

(2.1) *The morning star* is a body illuminated by the Sun.

(2.2) *The evening star* is a body illuminated by the Sun.

Both italicised noun phrases refer to the same entity in the world, the planet Venus, so either both are true or both are false. But if somebody's sense of the expression "morning star" differs from his sense of "evening star", he can think that sentence 2.1 is true, and sentence 2.2 is false. This leads us to a central problem: How can we be sure that two linguistic signs share the same reference? This question opens a Pandora's box of other problems. In order to ascertain whether two signs co-refer, we need to find find the referent of a sign, but how? And how can we be sure that a given linguistic sign refers at all—which expressions are "referring expressions"?

For example, in the literature on definite noun phrases, referential uses of definite NPs are clearly distinguished from attributive uses (Donnellan 1966). The NP "the murderer of Smith" is used referentially in sentence 2.3, because it points to the person who murdered Smith, and attributively in sentence 2.4 where it predicates of the butler that he has murdered Smith.

(2.3) The murderer of Smith used a blunt instrument.

(2.4) The butler is the murderer of Smith.

(2.5) The murderer of Smith must be crazy

The referential use is also called *de re*, because it is intended to point to a specific referent, while the attributive use, where any referent is possible that fits the definite description, has been termed *de dicto*. In sentence 2.5, the NP "the murderer of Smith" is attributive if we assume that the person who uttered the sentence does not know who the murderer of Smith is—but whoever did it, whoever we can predicate of that he murdered Smith, boy, must he be stark raving mad.

There are constraints on the entities that can be referred to, as well. Russell (1919/1993) states very clearly that it is only possible to refer to entities which exist in the real world. Phrases like "a unicorn" or "the present king of France" do not refer. Since Russell analyses the definite article "the" as the claim that there exists exactly one individual on which the description in the definite NP fits, a noun phrase such as "the present king of France" is simply false, because present-day France is a republic. Some semantic analyses of genericity also assume that it is

[1] I use the classical English translation here, because that provides me with the terminology used in English-language publications.

not possible to refer to generic NPs. For further discussion, see (Carlson 1991, Krifka, Pelletier, Carlson, ter Meulen, Link and Chierchia 1995).

But to what extent are these philosophical discussions relevant for the use of referring expressions in actual discourse? Consider the following three discourses:

(2.6) [The last unicorn]$_u$ was a very gentle animal. With the help of a sweet little girl, [it]$_u$ managed to defeat the evil sorcerer who had confined [its]$_u$ species to the sea. When [the unicorn]$_u$ finally left the little girl to rejoin [its]$_u$ family, the girl was really sad.

(2.7) [The murderer of Smith]$_m$ must be crazy. [He]$_m$ first lashed out at Smith repeatedly with a horse whip, then [the murderer]$_m$ severed Smith's hands and feet, and finally, [this danger to society]$_m$ kicked the victim down the cellar stairs, where [the vicious killer]$_m$ stabbed Smith exactly one hundred times with Smith's own Swiss army knife.[2]

(2.8) [The lion]$_l$ is a mighty hunter. [He]$_l$ lives in Africa. If a linguist comes too close to [him]$_l$ when [he]$_l$ is hungry, [the lion]$_l$ will eat that linguist with relish. And be careful—[lions]$_l$ will make no difference between functionalists and generativists.

In these three discourses, definite noun phrases and pronouns are used to refer back to an attributive definite NP (the murderer of Smith), to something which does not exist (the unicorn), and to a kind (lions). In the following section, we will investigate how these problems can be attacked with the help of a new concept, the concept of "discourse referent".[3]

2.1.2 Discourse Referents and Discourse Entities

Karttunen (1976) examines the question why definite NPs and pronouns can be used in such a potentially objectionable way. To this end, he proposes to focus the attention away from the referents of philosophy, which were designed to formalise the truth conditions of a sentence, and towards discourse referents, which are designed to model how people process discourse.

Karttunen assumes that in certain contexts, an indefinite NP establishes an entity in the discourse. If such an entity, which Karttunen calls *discourse referent*, has been successfully established, it is possible to refer back to that entity with a definite NP or a pronoun. This is the linguistic test he uses throughout his paper. Now, the initial question becomes: Under which conditions do indefinite NPs establish a discourse referent? He found that there are two main contexts in which discourse referents can be established:

1. The indefinite NP occurs in a sentence that "is asserted, implied, or presupposed to be true, and there are no higher quantifiers involved." (Karttunen 1976, page 383)

2. The indefinite NP occurs

[2] Note that I use the male personal pronoun here because that is commonly used in detective stories when the inspector does not have a clue who committed the crime.

[3] By jumping from Russell to Karttunen in this way, I naturally skip most of the very lively discussion of reference in the literature on philosophy and formal semantics. For a collection of classic papers, see e.g. (Moore 1993), and for recent summaries of the philosophical discussion, see e.g. (Newen 1995, von Heusinger 1997).

(a) in a sentence which reports on another world, the world of a person that is not the speaker. Example:

(2.9) John says that [an employee of Smith]$_m$ killed him. John thinks that [she]$_m$ did it because Smith kept making passes at anything with a skirt, and [she]$_m$ must have been fed up with this. These advances might have stirred up some old trauma in [her]$_m$, which would explain why [she]$_m$ butchered Smith this way.

(b) in a sentence which belongs to another "mode of discourse", such as counterfactuality. In such cases, the sentence in which the anaphor occurs has to continue in the same mode, or be evaluated in the same world as the sentence where the referent that this anaphor points back to was established. Example:

(2.10) When he grows up, John wants to marry [a rich girl]$_g$. [She]$_g$ should be old and terminally ill, so that he can inherit [her]$_g$ money soon. Needless to mention that [she]$_g$ need not be beautiful.

Karttunen (1976, page 383) calls such referents "short-time referents", because their life span is limited to the discourse segment which is in the same mode as the sentence where the referent was established.

To sum up, in discourse, referents may have a limited lifespan that is dictated by the conditions under which they are first mentioned, and it is even possible to refer back to expressions that have no referent, strictly speaking. We can cope with these unruly anaphors by introducing an intermediate level of representation for processing, be it mental or computational. This is the level on which discourse referents are established, maintained, accessed, and cease to exist.

After Karttunen, work on discourse referents has focused on the analysis and processing of referring expressions. The criterion for determining whether a new discourse referent has been created has largely remained the same that Karttunen proposed in his paper. Two of the first semanticists to integrate Karttunen's results in a formal semantic framework were Kamp (1981) and (Heim 1983). Kamp proposed a formalism, Discourse Representation Theory (DRT), that was intended to model how speakers dynamically update their interpretation of a discourse during communication. While Kamp's approach is rather abstract, Heim (1983) builds on a catchy metaphor: she sees discourse referents as file cards. On these cards, all information is entered which a discourse supplies about that referent.

Modern formal semantics deals with this problem in two ways. Firstly, reference is determined with respect to possible worlds, not only with respect to the world we live in. And in one of these possible worlds, there might well be a unicorn. Secondly, definite referring expressions are not resolved directly to individuals in a possible world, but they are interpreted via some sort of intermediate construction, such as discourse referents (Karttunen 1976), pegs (Groenendijk, Stokhof and Veltman 1996), or discourse subjects (Dekker 1998). The different proposals that have been made over the years in this tradition have centred more on exploring the nature of these discourse referents, and the ways in which they can be used to interpret referring expressions. The concept of entity status, in contrast, is geared to describing how a given discourse entity is introduced, accessed, and updated in linguistic communication, and how it relates to the discourse it occurs in. Entity status is a functional heuristic, not a formal construct, although parts of it could surely be formalised with a sufficiently rich formal apparatus. But this would

lead us too far afield here.

Discourse Entities

In a series of influential papers and theses in the late Seventies and early Eighties, Sidner and Webber proposed a different terminology in order to avoid confusing discourse referents with the concept of referent that is used in the philosophical discussion. They suggested that discourse referents be called "discourse entities", in order to avoid confusion with the referents in philosophy. These discourse entities are part of a *discourse model*, which represents the discourse as it has evolved so far. According to Sidner (1983), these entities are memory elements which contain a set of specifications. The only way to test whether a discourse entity has been established successfully is to test whether it can be referred back to with a definite NP or a pronoun. This test does not imply anything about the properties of that which is specified by the memory element. It can be a situation, an individual, an event, a kind, . . .—as long as it can be referred back to, it is a discourse entity. Referring expressions specify discourse entities. Two referring expressions that specify the same discourse entity are said to *co-specify*. A hearer can interpret a referring expression in two ways. If the expression specifies a memory element which is already in the discourse model of the speaker, but not in that of the hearer, the hearer has to add a new entity to his discourse model. If the discourse entity is already supposed to be in the hearer's discourse model, he needs to search for the entity in his discourse model which is specified by that expression. When a referring expression causes a new discourse entity to be introduced in the model of the current discourse, we will say that it *evokes* that entity. If the entity that an expression specifies already exists, the expression *accesses* that entity.

Discourse entities are the basis around which discourse models are organised (Webber 1981). They fulfil similar functions in Heim's and Kamp's systems of dynamic semantics. In contrast to Heim's File Change Semantics and Kamp's Discourse Representation Theory (DRT), Webber and Sidner do not develop a new semantic formalism. Instead, they use the traditional language of predicate logic and algorithmic pseudocode. In particular, Sidner's aim is to develop a conceptual framework that allows her to model anaphora resolution computationally. Using discourse entities, she can specify resolution algorithms on the basis of an abstract model that can be made computationally tractable. The following definition nicely summarises this computational viewpoint:

> A discourse entity is a variable or placeholder that allows us to index the information about an object or event that we extract from utterances to the appropriate mental representation of the object or event. (Jordan 2000, page 1)

Although (Kamp 1981), (Heim 1983), (Webber 1983), and (Sidner 1983) are frequently cited together, there are clear differences between the three strands. As evidenced by his textbook (Kamp and Reyle 1993), Kamp is by far the most cautious. He merely uses discourse markers as identifiers in the semantic formalism and, as far as I can see, refrains from any further claims. Heim (1983) goes one step further. She explicitly designs her file cards as a way to link Karttunen's discourse referents to referents in the real world and defines complex management operations on them. These operations have been developed further by Vallduvi (1990) in his analysis of information structure. Although file cards appear to be just the data structure

we need here, work after Heim has shown that the mechanism she proposed is not very flexible, because the underlying "card" metaphor is too static. As a consequence, even though File Card Semantics has influenced dynamic semantics to a great extent, very little work is couched in this framework today. Not so with DRT, which is still very much alive and is continuously modified and extended.

Webber (1983) and Sidner (1983) give the term a more computational interpretation. Both use discourse entities as tools for describing what anaphoric expressions can be resolved to. As far as I can see, neither of them has focused on constructing a semantic formalism around the concept of discourse entity; instead, the term has been used to describe and motivate solutions couched in the terms of first-order logic or algorithmic pseudocode. This makes the concept of "discourse entity" more flexible than that of a "file card" or a "discourse marker", which are associated with specific approaches to formal semantics. An additional advantage of the term is that "entity" appears to be a relatively colourless term, contrary to "referent" as in "discourse referent", which reminds its user of the problem we have been trying to escape, reference, or "subject" as in "discourse subject", which is homonymous with the grammatical subject. For these reasons, I will adopt the term "discourse entity" here.

The Semiotic Perspective

The introduction of discourse entities can also be justified from a semiotic point of view. Let us begin with traditional Saussurean semiotics. As linguistic signs, referring expressions consist of a *signifiant* and a *signifié*. According to de Saussure (1916/1985), defines these as two sides of the same coin. Both exist in the same sphere. The *signifié* of a referring expression is therefore not an object in the world, but a representation of an object. Discourse entities play a similar role. They live on an intermediate level between referring expressions and objects in real or possible worlds, and this level is a level of representation. In semiotics just as in computational linguistics, researchers have preferred to associate this level with the mental sphere. [4] If we want to emphasise that processing signs is inherently dynamic, we can also speak of signs as "mental constructions".

Discourse entities also have a place in Peircean semiotics. For Peirce, signs are triadic structures which connect an interpretant, which participates in the process of semiosis, to the corresponding *object* in the world. The *interpretant* is not the person who interprets the sign. Instead, it is that which makes the sign interpretable. Thus, in a Peircean framework, discourse entities would be the interpretants of referring expressions.

Eco (1994) develops Peirce's approach a step further. According to him, semiotics is concerned with signs as social forces. Whether a sentence in which a linguistic sign occurs is true or false, is outside the domain of semiotics. To claim that the referent of a linguistic expression is connected with something in the real world is counterproductive, because the meaning of a sign is crucially determined by social processes. Following Peirce, Eco conceives of the interpretant as something which can only be described by other signs. This has two consequences: Not only are signs always related to other signs, but the process of interpreting a sign is potentially infinite. This infinite process of semiosis means that

[4] Since many semioticists have interpreted de Saussure's *signifiés* as mental representations (Juchem 1984)

As far as I can see, this semiotic motivation has been completely overlooked in computational linguistics. It would certainly be interesting to study the connection between a semiotic interpretation of discourse entities and their cognitive or computational interpretation more closely, building on more recent work of Eco (2000), where he attempts to do the same for concepts (c.f. in particular his Chapter 3). But to explore this avenue further is beyond the scope of this thesis. First it is about entity *status*, not about the nature of discourse entities, second, entity status is *per definitionem* a very prosaic notion, designed to help describe how discourse entities manage to fit in with their co-text and the discourse model.

2.1.3 Entity Status

What could the status of a discourse entity be? Remember that we introduced entity status as an umbrella term which groups together properties of discourse entities that influence the extent to which a discourse entity is given. But since givenness itself is an ill-defined notion, we need to find another criterion for deciding what should go under that umbrella. The solution is to look at the questions which researchers have tried to answer by recurring to something they then called "givenness": How can discourse entities be accessed? How are they established and initialised? How central is a discourse entity to the discourse it occurs in? Hence, my definition of entity status is fundamentally functional because the properties we will assemble should help us understand how people use discourse entities to process and produce language. I will also strive to describe these properties in a way that is as theory-neutral as possible, because once we have a functional definition of the properties that we might need, we can use this as a checklist to select an adequate basis for formalisation.

The research questions mentioned in the last paragraph split naturally into two dimensions:

The dimension of structure. Key desideratum:

> *The status of an entity contains information about how central it is in the discourse.* This information is multi-faceted: What is the relation between that entity and the general topic of the discourse? Are there discourse segments in which it is more or less central? How has its status developed so far?
>
> If we accept the premise that discourse models are built around discourse entities, then the more central an entity is to the discourse (or to some segments of the discourse), the more important it is for grouping (*structuring*) the propositions in the discourse.

The dimension of management. Key desideratum:

> *The status of an entity contains information that is necessary for managing it.*
> As we will see in Chapter 4, three types of operations are needed for this purpose: initialisation, access, and update. The management dimension supplies the information that is needed for these operations, but only to the extent that it can be derived from the referring expressions or that it is already encoded in the discourse model.

Although both dimensions are related, they are evidently not the same. The first dimension cannot be defined without a theory of discourse structure, and the second dimension cannot be defined profitably without a theory of memory.

The second dimension, on the other hand, is the dimension of *management*. Whenever the hearer hears a new referring expression, he has to know where he can find the specifications he needs for interpreting it.

As we will see in Chapter 4, the management dimension has been by far the most popular in the literature, and researchers have identified several sub-dimensions. The structure dimension, on the other hand, is related to a variety of theories: theories of coherence, theories of discourse structure, and theories of topic. These relations will be discussed further in Chapter 3.

At this stage, we know that discourse entities are conceptual tools for describing discourse, that they are supposed to be the elements around which discourse models are organised and which can be referred back to anaphorically, and that their status has a management and a structure dimension. But can this useful heuristic tool for (computational) linguists be integrated into a theory of human communication? Can we successfully transfer the notion of discourse entity to a non-algorithmic level of thinking about communication? That is the question that I will examine in the following section.

2.2 Discourse Entities in Communication

Discourse entities in their many guises have proved to be an important conceptual tool for linguistic analysis, formal semantics, and computational linguistics. We will now take a step back and look at the concept from a more general perspective, the perspective of communication theory. As I have noted in the introduction to this chapter, the perspective on discourse entities and their status that I will develop in this section serves two purposes for the argument I wish to develop:

1. to show that there is more to discourse entities than a purely computational approach suggests (Section 2.2.1)

2. to find out how aspects of discourse entities and of their status in discourse could be investigated quantitatively (Section 2.2.2)

For those readers that are not familiar with Gerold Ungeheuer's approach to communication theory, I sincerely recommend a detour via Appendix D.

2.2.1 The Communicative Perspective

When we apply Ungeheuer's approach to linguistic problems, we face a fundamental problem: In linguistics, we strive to detect regularities in the way that linguistic signs are used, and, if possible, these regularities should be described in the framework of a rigorous theory, so that we can predict further regularities from them. But for this purpose, we need to abstract away from the situations in which the signs are used. Before we can start to analyse any communicative actions, even our own, we need to observe what is being done. As soon as we start to observe, we switch to the external perspective, and lose the immediate access that only the internal perspective yields. Ungeheuer reflects this problem in the dichotomy *communicative/extra-communicative* ("kommunikativ / extrakommunikativ"). Ungeheuer introduces the distinction between communicative and extra-communicative in (Ungeheuer 1970/1972b), where he discusses the reality of phonemes. From an extra-communicative point of view, phonemes are

structural units that make up words and that can be identified by rigorous analysis of observed data, following e.g. the procedures of (Trubetzkoy 1939/1989). From a communicative point of view, people hear the mesh of acoustic cues that is an utterance and make sense of it given the situation they are in and the experiences they have had so far.

In other words, when we use language communicatively, we perform acts of communication with it. When we observe how language is used in communication, we switch to the *extra-communicative* perspective—we are "outside". In fact, this distinction follows directly from the fact that verbal communication is an action, and that actions can be characterised from both an internal and an external perspective. The internal perspective is the communicative one, the external the extra-communicative.

By definition, linguistics is extra-communicative. The extra-communicative perspective is that of the analyst who wishes to find stable generalisations that hold across many different situations, the communicative perspective is that of the individual who uses language in a given situation. The important point here is that just because we find that some structural units work when we analyse observed data, this does not mean that people really use them. Both perspectives are equally necessary, and both perspectives complement each other, but they should not be mixed. Although Ungeheuer admits that the perspectives are related, he advocates that they should be strictly separated, so that the results of one are not contaminated by the other.

> Beide Betrachtungsweisen, die kommunikative, und die extrakommunikative, sind in der Phonetik gleich wichtig, ihre kategorialen Unterschiede müssen jedoch mit aller Klarheit festgehalten werden. Die Ergebnisse aus den beiden Forschungsgebieten stehen nicht—wer würde dies auch vermuten?—beziehungslos nebeneinander; die Beziehungen können aber erst adäquat analysiert werden, wenn der wissenschaftliche Wert dieser Ergebnisse nicht durch Kontamination kommunikativer und extrakommunikativer Gesichtspunkte herabgesetzt wird.[5]
>
> (Ungeheuer 1970/1972b, page 82/page 46)

But where does contamination begin, and where does productive dialogue end? The answer to this question is given in the first sentence of the quoted passage. Ungeheuer clearly states that both perspectives are necessary. Why? There can only be one reason: There must be some questions that can only be answered from one perspective, but not the other, and vice versa. This is the principal area of dialogue between the two perspectives. Extra-communicative analysis can discover highly interesting structures, and put these structures to use in applications. While an extra-communicative approach is eminently suitable for analysing external actions that are observed in several situations, only communicative analysis can venture informed guesses about the internal actions as well, can hope to understand why the process of communication developed in a specific way in a given situation. But not only do the two approaches, the communicative and extra-communicative one, complement each other. They also share a common basis of phenomena which they investigate.

[5] *Both points of view, the communicative and the extracommunicative, are equally important in phonetics, but their categorical differences need to be stated as clearly as possible. The results from both fields of research do not stand stand beside each other unrelated—who would assume that?—but these relations can only be analysed adequately if the scientific value of these results is not contaminated by mixing communicative and extra-communicative aspects.*

Let us now turn back to the topic of this section, "entity status in communication". So far, we have only dealt with an extra-communicative perspective on entity status, the perspective of (computational) linguistics, which was discussed in Section 2.1.3. This will also be the main perspective of the rest of this thesis: I will analyse text corpora quantitatively, and I will explore probabilistic models of entity status; and in doing this, I will have to largely abstract away from the communication situation in which the texts were originally produced and focus on the systematic regularities in linguistic behaviour that we can detect using methods from statistics. What am I losing there? What is the communicative perspective?

In order to answer this question, we need to go back to the question which has motivated computational linguists to introduce discourse entities as a unit of analysis: How come we can refer back to something anaphorically? Anaphoric referring expressions are linguistic forms. As such, they cannot be part of the primary content. From a communication semantic point of view, hearers interpret them according to two sets of rules:

1. linguistic–semantic rules: Every language has a set of conventions for determining the meaning of sequences of linguistic symbols. The speaker uses these conventions for planning the external action by which she seeks to influence the hearer, and the hearer uses these conventions during the internal actions that he performs in order to interpret the behaviour that he perceived the speaker produce.

2. rules of communicative interaction: Speaker and hearer follow these rules in order to communicate as successfully as possible, in order to understand each other.

When anaphoric referring expressions occur in a sequence of symbols, it is usually important to resolve the pointer correctly if the communicative act is to succeed. The hearer needs conventions for resolving anaphora, if he wants to successfully re-construct the primary content of the speaker's sequence of signs, especially its material component. But according to which conventions should the speaker choose an anaphoric expression, and which conventions should the hearer use when he encounters it? This problem belongs to the modal component of the primary content. As we have seen in Appendix D, the modal component modifies a given material component in a variety of ways. In the context of anaphora, one level of the modal component is of special interest: modifications with respect to the communication situation and the other participants in the communication process. For example, suppose you are the Republican presidential candidate in the 2000 U.S. Campaign, and you are lagging behind your opponent in the polls. You know that you need good press coverage, journalists that extol your virtues, that forget your slips in speeches. In this situation, you would only call a journalist a "primary league ass-hole" when you can rest assured that no journalist is listening in. You would not dare to do so if the microphone you have been speaking into were still on.

So far, so good—I have merely paraphrased old linguistic insights from a slightly different perspective. To make the picture complete, one element is still missing: that which the anaphoric expression refers back to, that which it picks out. In the extra-communicative perspective, this is the discourse entity. But what is it in the communicative perspective? Anaphoric referring expressions clearly pick out something that is experienced by both speaker and hearer as something that they experience as an "unit" at the time that it is referred to in conversation, as something that they can predicate something of, something that they can associate experiences with, something that is part of their personal experience theory (PET), the sum of all

experiences they have ever made. Can we be more precise than this? No, and you do not need to look at the sweeping definition of potential discourse entities that is given by e.g. (Webber 1991) to see that. Just perform a simple *gedankenexperiment*. Search your memory, which has been formed by your system of experiences, and try to find an expression that summarises a single experience, a set of experiences, an even larger set . . . Once you have constructed a set of experiences as a unit, and once you can give that unit a name, you have constructed something very much like a "discourse entity". This unit that you have just constructed is not static. It constantly changes shape during the communication process. Such a unit can fade and be revived, or it can become a fixed part of the way that the PET is organised. Take the example of "German national soccer team". Somebody who neither knows nor cares about soccer and who happens to hear an item about the team's dismal performance of the radio will briefly construct a unit which helps him understand that there is a team sport which people call soccer, that Germany has a national team, and that this team has had a few problems with winning matches in a Europe-wide tournament recently that they call "European Championship". As soon as the radio item is over, the soccer–ignorant will forget about it. On the contrary, for a German soccer fanatic, many experiences are associated with the soccer team: good games, bad games, coaches, players, or feelings of joy when the team won its three World Championships. This person instantly connects what he hears to his experiences, especially to his recent experiences of pain and his memories of dismal players. A unit such as "German national soccer team" has no sharp boundaries. It is connected to the system of your experiences. Depending on the direction from which you access it, depending on the experiences you have made immediately before, depending on the experiences which you have recalled or been vaguely reminded of, different aspects will be more or less readily accessible to you.

How speaker and hearer deal with the differences in their respective PETs, how they negotiate the common ground on which they can build their conversation, and whether they can build such a common ground at all, that has been the focus of many experimental studies in psycholinguistics. These studies are usually designed around a heavily constrained task: describe tangram figures (Clark and Wilkes-Gibbs 1990), guide each other around a map (Anderson et al. 1991) , describe video films, and so on. For an in-depth discussion of this methodology, see (Brown 1995).

Clark and Haviland (1977) proposed that there is a *given–new contract* between speaker and hearer. The speaker presents the given information first, so that it is easier for the hearer to embed the new information into his discourse model. In other words, if speaker wants to make it easy for her hearer to interpret what she has just heard in terms of her PET, she will try to make these connections explicit, and she will verbalise these connections as early in the message as possible, since they will help her hearer process what follows. But what is this given information? Clark and Marshall (1981) rephrase this question as: Which knowledge do speaker and hearer need to share, and how can they discover they share it? If the speaker not only wants to know what the hearer knows about the subject they are talking about, but also what he assumes about her, and what he assumes that she assumes about him, and what he assumes that she assumes that he assumes that she assumes about him, we are soon stuck in a nice infinite loop. Ungeheuer would have probably called a halt right after the first recursion, because he argues that it is per se impossible to know another person's PET. Clark and Marshall (1981) suggest that in ordinary conversation, speakers and hearers cope with this problem by a set of heuristics. When processing a definite description, speakers and hearers assess where

Knowledge source	Example and explanation
COMMUNITY MEMBERSHIP	The luggage (known to Discworld cognoscenti) `http://www.lspace.org/books/whos-who/luggage.html`
CO-PRESENCE	
physical	The luggage is in a hotel room. Rincewind asks Twoflower: "And the luggage really washes and irons?"
immediate	both are looking at the luggage
potential	both are about to enter the hotel room and see the luggage
prior	both have just left the hotel room and seen the luggage
linguistic	
potential	When Rincewind packed [it]$_L$, [the luggage]$_L$ grunted. (*cataphoric*)
prior	When Rincewind leaves without [the luggage]$_L$, [it]$_L$ is upset. (*anaphoric*)

Table 2.1. Sources of mutual knowledge. In case you wondered, Discworld Luggages are wooden chests on legs.

they might know the discourse entity that the description accesses from. Clark and Marshall (1981) name four sources of mutual knowledge, which community membership, physical co-presence, linguistic co-presence, and indirect co-presence, which is a mixture of community membership and physical or linguistic co-presence. These sources are evaluated by a set of heuristics. Table 2.1 gives examples for the first three sources of mutual knowledge.[6] From the communication theoretic standpoint that we have taken, we should not be surprised that Clark and Marshall (1981) managed not more than just that, a set of heuristics. Since speakers can never really know the PET of their hearers, they are by the very nature of the communication process confined to assumptions and heuristics.

When the speaker wants to make sure that the hearer accesses the right discourse entity, she has many different strategies at her disposal. But not all speakers are equally cooperative, or adept at finding the right strategies. When such failures occur, then that is too bad for the conversation, but excellent for the analyst, who can now get to work and unearth the reason for these failures (Brown 1995). The results of Bard, Anderson, Sotillo, Aylett, Doherty-Sneddon and Newlands (2000) on Map Task data indicate that in fact, speakers are more egocentric than research has assumed so far. Whether the referring expressions speakers produced were more or less intelligible not only depended on whether the discourse entity was new to the hearer, but also on whether it was new to the speaker. Keysar (1997) argues that analysts should be cautious with notions such as common ground or mutual knowledge, which are difficult to model, and only assume them where they are needed to explain certain kinds of cooperative behaviour. But differences in referential strategies need not only be due to differences in cooperativity. Other factors are age (Light, Capps, Singh and Albertson Owens 1994, Clancy 1992, Vion and

[6] We will meet similar taxonomies of knowledge sources again in Chapter 4.3, when we discuss how linguists have classified referring expressions.

Colas 1999), social class (e.g. Hemphill 1989, and the references therein), and language/culture (Clancy 1980, Pu 1995).

2.2.2 Methodological Consequences

We have seen that from a communicative point of view, discourse entities are dynamically changing mental units which are more or less fixed, depending on how deeply they are anchored in a person's system of experiences, her PET. We have also seen that it is in principle impossible to model the PET of somebody else exhaustively. Hence, all categories that rely largely on the PET, or, to put it in a less philosophical way, on the mental states of speaker and hearer, are problematic. These problems surface sharply when these categories need to be used for annotating corpora. In such situations, we find two main types of problems:

1. The annotator lacks knowledge about the communication situation. Often, she knows little about the particular context in which a discourse was produced, and of course, she cannot peek into the minds of speaker and hearer as they produce and process the linguistic signs she is annotating. Although psycholinguistics has developed sophisticated experimental setups in order to address this problem, these time-intensive techniques are usually not available to corpus annotators, who need to rely on their own intuition as language users and on their world knowledge. As a consequence, the categories to be annotated should be as independent from the communication situation as possible. Recall that we are dealing with observed language, and that we want to find and describe patterns in these observations. The more we can reduce the communication situation to a few parameters, and the more context we can exclude, the less do we have to infer and to guess, and the more can we reliably observe. For discourse entities, something that can be readily observed is co-specification. If an annotator cannot determine co-specification sequences correctly, it is likely that she does not understand the discourse. In our work on the BROWN-COSPEC-corpus, documented in Appendix C, we encountered this problem a few times with argumentative scholarly texts which differentiated e.g. between the "concept of nationalism" and "nationalism". Another problem occurs in literary texts when an author deliberately leaves the identity of some people or things unclear or when a protagonist does not know that two persons are identical. To solve these problems, we would need to index co-specification relations with the contexts in which they hold (Wiebe 1991). Still, correct co-specification relations are so fundamental to building an adequate mental representation of a text, that we can safely expect annotators that know enough about the subject matter to mark such sequences reliably.

2. The annotator can only assign the annotation categories on the basis of her PET. This trivial statement accounts for a few interesting observations. For example, Teufel (1999) reports that although her first annotation scheme for research articles was stable, i.e. she achieved a high degree of agreement with herself when she labelled the texts again after four weeks, the scheme was not reliable: two other annotators who were not specialised in discourse analysis did not match her annotations well. The more an annotation scheme makes use of categories that allow for a wide range of slightly different interpretations, the less reliable it is bound to be.

In the context of discourse analysis, Mann, Matthiessen and Thompson (1992) note that inter-annotator agreement for text annotations based on Rhetorical Structure Theory is low. It is

interesting that both schemes, Teufel's original scheme and RST, heavily operate with categories that describe how readers interpret a part of discourse, what they construct that part of discourse to be. The following quote nicely illustrates this with categories from a set defined in (Teufel 1998):

> To give an example, we were not sure about the right annotation for the following sentence: *We then show how different classes of pragmatic inferences can be captured using this formalism, and how our algorithm computes the expected results for a representative class of pragmatic inferences.* (S-29,9504017)
> Is the sentence to be counted as TOPIC, because "*pragmatic inferences*" are the TOPIC of the paper? Or is it the case that "*capturing different classes of pragmatic inferences*" is the PROBLEM/PURPOSE? Or should this sentence be classified as SOLUTION, as the phrase "*our algorithm computes the expected results*" could be interpreted as a high level description of the approach used? (italics in the original, Teufel 1999, page 108)

Is it bad if annotators cannot agree on annotations? Not really. If the annotation strategies of the annotators can be proved to be consistent, then the annotations reflect reasonably stable categories of the annotators' PET. Hence, their annotations actually provide two kinds of data: data about how linguistic forms can be categorised, and data about inter-individual differences in the boundaries of these categories. But stability is a necessary precondition for this type of analysis.

I do not deny that reliable annotation schemes are very valuable for linguistic research. Once a scheme has proved to be reliable, many annotators can collaborate in annotating large amounts of text, which can then be collected into a large, homogeneous corpus. Reliability is also crucial if we want to establish "gold standard" data sets for certain computational linguistic tasks, such as Word Sense Disambiguation (Kilgariff 1998), co-reference annotation (Hirschman and Chinchor 1997, Hirschman, Robinson, Burger and Vilain 1998), and document summarisation (Teufel, Carletta and Moens 1999). But that should not lead to cheap polemic against those researchers who analyse discourses in depth, trying to guess at the motivations of speakers and hearers. As Brown (1995) has pointed out, such guesses are invariably coloured by the analyst's PET (although she did not use this term). However, this need not be detrimental, as long as the analyst is acutely aware and scrupulously honest about these necessary limitations. In many sociolinguistically oriented studies such as (Tannen 1979, Selting 1988), the researchers interview the people who were recorded carefully in order to check their intuitions.

2.3 Summary

Along with many scholars from the humanities, and some from the sciences, I assume that it is not possible to develop a comprehensive, mathematically rigorous formal model of communication. But I also assume that formal methods are necessary for describing the patterns and structures that can be observed in communication in a precise and succinct way. Since most of the results that I will present in the later chapters of this thesis are quantitative, statistical models and statistical analyses of observed data, I need a perspective that shows very clearly

the *limits* of such a computational approach, an approach that highlights what we lose when we start counting and stop interpreting.

This is why I took a somewhat unusual perspective in this chapter, the perspective of semiotics and communication theory. From the point of view of Peircean semiotics, it is perfectly natural to distinguish between a *referent* that links a sign (here: a referring expression) to an object in the world, and a discourse entity that connects that sign to other signs in the process of semiosis. In Peircean terms, a discourse entity is nothing but the *interpretant* of a referring expression. This is an interesting semiotic twist on the usual explanation of discourse entities as mental representations of what anaphoric expressions can refer back to (Sidner 1983).

From a communicative point of view, discourse entities are an "extra-communicative" (Ungeheuer 1970/1972b) formalisation of units in the flow of experiences that speakers and hearers have in discourse. The boundaries of these units are fuzzy, and they are enmeshed in a web of experiences which constitutes each person's personal experience theory (PET). The units are dynamic: each time they are used in understanding discourse, the net of experiences of which they consist changes ever so slightly.

The status of a discourse entity can be defined along two dimensions:

the structure dimension, which states which role an entity plays in the stretch of discourse it occurs in, and

the management dimension, which provides information that is needed when a discourse entity is initialised, accessed, and updated.

Both dimensions overlap: the more central an entity is in the discourse, the easier it is to access that entity. And both dimensions can be labelled with the term "givenness": an entity which is central to the discourse and which can be accessed without any problems is "given" to both speaker and hearer, and an entity which has yet to be integrated into the discourse model is "new". In Chapter 3, I focus on the structural dimension—I will relate the concept of entity status to other models and concepts that are used for analysing discourse: discourse structure, topic, and relevance. The next Chapter, 4, relates the management dimension to previous concepts in the literature.

3 The Dimension of Structure

In this chapter, we will discover what lies behind the "structure dimension" of entity status that was introduced in Chapter 2. The structural status of a discourse entity covers how important that entity is for the discourse—both for its content and its structure. In the first section, 3.1, I describe what a theory of structural entity status needs to explicate. The following sections relate previous theoretical work on discourse to structural entity status. I begin in Section 3.2 with the most fundamental notion, coherence. Then, I move on to the key theories for explicating structural entity status, theories of text and discourse structure (Section 3.3). From the wealth of proposals, I select three that are important in psycholinguistics and computational linguistics, Rhetorical Structure Theory (Mann et al. 1992), Grosz and Sidner's theory (Grosz and Sidner 1986), and van Dijk's theory (van Dijk 1980). For each of these theories, I explain how structural entity status could be explicated in its framework. Finally, I move to a phenomenon that is intimately connected with the role of a discourse entity in the discourse it occurs in: topicality (Section 3.4). Section 3.5 summarises the main conclusions.

3.1 What is Structural Entity Status?

Just as the concept of entity status itself, structural entity status is a convenient conceptual shortcut for a whole bag of related information:

1. In which discourse segments does the entity occur?

2. How is the entity linked to other entities in the discourse?

3. How closely is the entity connected to the speaker's communicative intentions?

The relation between these three questions, the overarching question, is simply: What role does the discourse entity play in the discourse where it is mentioned? Contrary to established practice in computational linguistics, these questions are not phrased in terms of mental or algorithmic representations. Instead, the questions refer to the discourse itself and to the situation in which that discourse was produced. Their wording only assumes that the discourse we are analyzing can be described at some arbitrary level of structure, and that we are looking at an instance of communication.[1] Let us now consider each of these questions in more detail.

[1] Remember that, contra (Watzlawick, Beavin and Jackson 1967), and with Ungeheuer, I do not think that it is impossible not to communicate. For a resolution of this triple negation, see Appendix D.

Position of the Entity: The simplest way to answer this question is to protocol where the entity is mentioned in the text, in other words, to establish co-specification sequences. As I will argue in Chapter 5.4, this is in fact the best corpus-based operationalisation of entity status that we can get, if we want reliable results that can be processed by statistical methods of analysis.

Quite a few researchers have argued that the threads that co-specification sequences and similar sequences of substitutions weave through discourse are what keeps discourse together. A classic example is (Harweg 1979). Harweg went as far as claiming that the syntagmatic relationships between pronominal forms and the expressions that they substitute sufficiently characterise a well-formed text. True, his interpretation of the term "pronominal forms" is very wide; he used the term to cover many kinds of anaphoric relationships. But his approach can be criticised from a more fundamental point of view: Why should text linguistics define well-formed texts structurally? More recently, Klein and von Stutterheim (1992) have presented an approach to text structure that is based on *referential movement*. If coherence is in the mind of the reader, if it crucially depends on his abilities to make semantic and pragmatic sense of the locutionary and illocutionary acts that are realised in a given text, as e.g. Brandt and Rosengren (1992) or Nussbaumer (1991) argue, then texts cannot be defined purely via their structure. This referential movement, however, is conceived of very broadly. It covers not only relations between anaphors and their sponsors, but also the movement through temporal and spatial coordinate spaces. In this approach, anaphoric relationships are still central, but the term "anaphora" is given a very broad interpretation.

Although referring expressions themselves are not as crucial as structuralist proposals might imply, they are still popular toys for researchers in discourse, because they provide both good examples and good diagnostics for local and global coherence. This attitude is expressed nicely in the following quotes:

> Just as linguistic devices affect structure, so the discourse segmentation affects the interpretation of linguistic expressions in a discourse. Referring expressions provide the primary example of this effect. (Grosz and Sidner 1986, page 178)

> Eventually, understanding the structuring of discourse should allow us to account for reference phenomena in texts.
>
> (Polanyi 1988)

Experiments and theory both point to a close connection between referring expressions and discourse structure. The results of e.g. Marslen-Wilson, Levy and Komisarjevsky Tyler (1982) or Vonk, Hustinx and Simons (1992) demonstrate that speakers and hearers regard definite descriptions as cues to episode boundaries when they occur instead of pronouns. The mental representation of narratives is oriented towards the central characters that occur in the narrative (c.f. the overview in Sanford and Garrod 1994). But verifying that connection on a corpus is an onerous task. Segmenting a discourse into its parts and determining the hierarchical structure of these parts is not trivial (Passonneau and Litman 1993). Although there are a number of important surface cues other than referring expressions, such as intonation (Thorsen 1985, Hirschberg and Grosz 1992) or cue phrases (Passonneau and Litman 1997), segmentation still depends to a large degree on the interpretation of the annotators. This is well documented by studies such as (Passonneau 1998).

Links to other entities: Such links can be established on many levels. The most traditional level is surely collocation analysis. But once we have a theory of the domain that our texts come from, we can define schemata and scenarios, and on the basis of this knowledge, and we can predict which discourse entities will co-occur and which bridges listeners will be able to build. Such links are very important for the management of discourse entities, as we will see later in Chapter 4. Discourse entities that are connected tightly to a person's world knowledge activate a web of related knowledge when they get anchored to the discourse model. The more entities from a common semantic field or scenario are evoked, the more other concepts from that field are activated, and the easier it becomes for the hearer to accomodate new discourse entities from the same topic area.

Connection to Communicative Intentions and "Gist": Whenever linguists discuss what a discourse can be said to be about, they use one or both of the terms theme and topic. Clearly, when we want to ascertain how important discourse entities are for describing the gist of a discourse, we need to determine whether they have been topical. Both terms have been filled with content in a bewildering variety of ways, and both terms have tended to be restricted to noun phrases. Given that discourse entities can be not only things and persons, but also situations, actions, concepts, or even whole scripts, this form-based fixation seems a little strange, although it is practical in a language such as English, where most discourse entities are indeed realised as NPs. In Section 3.4, I will attempt to shed some light on the issue.

Outlook: In sum, a theory of structural entity status, that is, a theory that explains which role discourse entities play in structuring and processing discourse, needs to integrate three contentious concepts, that of topic, that of coherence, and that of discourse structure, in a single framework. As yet, I have not come across such a framework. Therefore, aspects of structural entity status will be considered from the point of view of several different theories here, all of which have proved to be useful for parts of the (computational) linguistic community.

In Section 3.2, I will survey critically the connection between coherence and co-specification. The conclusion will be that since coherence is constructed largely in the minds of speaker and hearer, it relies on a multitude of cues, of which co-specification is but one, albeit an important one.

Similarly, discourse structure is to a large extent not observable. Despite these difficulties, various models of it have been proposed. In Section 3.3, I will relate structural entity status to three of the most influential ones: van Dijk's (1980) theory, which has greatly influenced psycholinguistic approaches to discourse, Grosz and Sidner's (1986) theory, which is the basis for Centering Theory (Grosz, Joshi and Weinstein 1995), Rhetorical Structure Theory (Mann and Thompson 1988, Mann et al. 1992), which emerged from Systemic Functional Grammar (Halliday 1994).

Finally, the more central an entity is to the discourse, the more justified it is to say that the discourse is in some sense "about" that entity—which leads us straight to the time-honoured mess of a concept that linguists call "topic". Its relation to structural entity status is discussed in Section 3.4. We will see that structural entity status provides an interesting link between sentence-level and discourse-level topics. The main results of this chapter are summarised in Section 3.5.

3.2 Coherence

Every time a discourse entity is mentioned in a discourse, it should be realised by an adequate referring expression; else, the resulting discourse will be incoherent. This is a central tenet of many scholars who work on the form of referring expressions in discourse. But is this role that coherence is supposed to play justified? Is coherence more than just referential continuity? What is coherence, anyway? Is it just an empty folk-linguistic term (Reboul 1997), or can it be useful in linguistic research? To me, these questions are so important that they warrant a small detour through linguistic (Section 3.2.1) and psycholinguistic (Section 3.2.2) theories of texture and coherence. An executive summary is given in Section 3.2.3.

3.2.1 Linguistic Approaches to Coherence

Almost all researchers agree that coherence is what makes an utterance or a sequence of utterances into a text. Although grammar, morphology, lexicon, and prosody provide means for signalling coherence, linguists agree that nonsense texts which show all necessary surface cues to coherence simply do not cohere—or if they do, they cohere very loosely. Therefore, coherence has to do something with the meaning of a text. But what coherence exactly is remains elusive. In scholars' theories, this concept wields considerable influence, especially in theories of anaphoric reference: justification enough to take a peek behind the veil of that fundamental yet fuzzy concept.

Some researchers appear to make coherence the main criterion for textuality (Halliday and Hasan 1976, Schade, Langer, Rutz and Sichelschmidt 1991). But such a perspective obscures more than it illuminates. No definition of textuality shows this more clearly than the seven criteria posited by de Beaugrande and Dressler (1981):

cohesion: The surface elements of a text should be connected by linguistic means, such as anaphora, sentence connectives, and so on.

coherence: The concepts and relations between concepts that occur in a text should be connected to each other and they should be relevant for each other's interpretation.

intentionality: A text should be produced by the communicator with a certain intention in mind.

acceptability: A text should be relevant to the addressee. If the addressee cannot process the text adequately, then that text is not acceptable.

informativity: A text is supposed to be informative. De Beaugrande and Dressler's concept of informativity relies on probabilities and classical information theory (Cover and Thomas 1991): the more informative, the less known or expected, the less likely to occur in the present context.

situationality: A text should be relevant to the situation it occurs in. For example, traffic sign messages need to be highly elliptical, lacking cohesion, but since this is completely adequate given that they need to be processed quickly, they can still be read as texts. If the

communicator uses the text to make aspects of the participants' model of the communication situation explicit, she is *monitoring* the situation, if she uses it to change the situation in a certain way, she is *managing* it.

intertextuality: A text is interpreted relative to other texts. This results in clusters of texts which adhere to similar textual conventions, *text types*. These conventions manifest themselves in norms and expectations about how to produce and process texts.

Although that catalogue of seven criteria can be criticised on several counts (Heinemann and Viehweger 1991, Brinker 1997), it makes an extremely useful point: that an addressee has managed to find connections between the meaning of the sentences of a text, that he has managed to discover coherence relations that connect parts of the text to others, does not imply that he has successfully understood that text. Take for example the utterance

(3.1) The window is open.

(3.2) It is freezing outside,

(3.3) and the cold begins to seep into this room.

The text is perfectly coherent. The sentence formed by (3.2) and (3.3) specifies the effect of (3.1), and (3.2) gives the reason for (3.3). But if the addressee of that well-formed, coherent discourse is sitting smugly in his armchair in front of the blazing fire, while the communicator stands by the window shivering, he obviously has not understood the message at all. Another good example where pragmatic conventions allow us to establish coherence is the following road sign example from (de Beaugrande and Dressler 1981, page 9, example 1):

(3.4) Slow
Children
At Play

For Halliday and Hasan (1976), coherence is the main criterion of textuality. That a text coheres is signalled linguistically by various means. Halliday and Hasan (1976) call those means *cohesive* which connect parts of the text to each other: reference, which subsumes our co-specification, substitution, ellipsis, conjunction, and lexical cohesion, which includes repetitions and collocations. More generally, cohesion is a relation between two passages of text A, B, where A is necessary for the interpretation of B. In the text itself, cohesion is a process, because the cohesive relations in a text are interpreted while the addressee interprets the text.

Since Halliday and Hasan aim to explore language in use, they cannot define coherence solely in terms of the co-text, they also need to take into account the context in which a discourse is produced and processed.

> The concept of COHESION can therefore be usefully supplemented by that of REGISTER, since the two together effectively define a TEXT: A text is a passage of discourse which is coherent in these two regards: it is coherent with respect to the context of situation, and therefore consistent in register; and it is coherent with respect to itself, and therefore cohesive. Neither of these two conditions is sufficient without the other, nor does the one by necessity entail the other.
>
> (Halliday and Hasan 1976, page 23; emphasis in the original)

A register is a set of linguistic features which tends to be used in certain communication situations. These situations are not only characterised by the participants, the social relations between the participants, the intention that the communicator pursues with the text and the events in which the production of the text is embedded, but also by the *context of culture* (Firth 1950, Malinowski 1923). It is interesting that most scholars who quote Halliday and Hasan on *coherence* do not even mention register (and then criticise them for failing to explain why cohesive nonsense texts fail to cohere). But it is easy to understand why. Register has a quite specific meaning here. The term is defined using three basic, complex notions of systemic functional theory, field, mode, and tenor. The definition of these concepts is in turn rather abstract and metaphorical, which is not unusual in systemic-functional linguistics. No wonder Halliday is difficult to understand without some grounding in systemic-functional linguistics and its philosophy. The main problem with the approach as outlined in (Halliday and Hasan 1976) is the focus on *local* coherence phenomena. But then again, many systemic functionalists are working on text structure, and at least one major theory of discourse structure, Rhetorical Structure Theory (RST, Mann et al. 1992) is rooted in SFL. Thus, if we put the work of Halliday and Hasan in its context, the criticism is again groundless.

In the systemic-functional approach, co-specification is part of one of five factors contributing to cohesion, and cohesion is one of the two aspects of coherence. Although Halliday and Hasan differ from de Beaugrande and Dressler in that they define coherence as their main criterion for textuality, a closer comparison of the criteria of both shows that the two definitions actually overlap to a large degree. Intentionality, informativity, intertextuality, situationality, and acceptability are taken care of by register. The cohesion of Halliday and Hasan subsumes cohesion, and, to a certain degree, the coherence of de Beaugrande and Dressler—if the concepts that a text evokes cannot be related in any way, it is not possible to relate some parts of the text to others semantically.

The discussion of both (de Beaugrande and Dressler 1981) and (Halliday and Hasan 1976) points to the conclusion that coherence is really a pragmatic notion: A text is coherent if the addressee can construct a connection between its parts, and the easier that connection is to build, the more coherent a text is. Fritz (1982) pursues this idea one step further. He views coherence relations as connections between verbal actions ("sprachliche Handlungen"). In order to establish these connections, we need linguistic cues, an understanding of the meaning of these sentences, and knowledge of social conventions that govern sequences of verbal actions. Essentially, this view of coherence does not differ greatly from (Halliday and Hasan 1976), but there are two important distinctions: Firstly, Fritz does not need the apparatus of systemic-functional grammar to state his conclusions, and secondly, his reformulation in terms of verbal actions shifts the main attention from the analysis of written text to the analysis of spoken language. It also provides a more convenient interface to sociolinguistics.

Hobbs (1979) describes coherence from the perspective of planning and goals. This perspective is very valuable for discourse generation, since it helps describe why the communicator chose which coherence relations between text passages. Hobbs distinguishes three levels of planning: the deepest level, where goals for communication are set and divided into subgoals, the level of coherence, where the communicator structures her message so that it will achieve the desired effects, and the sentence level, where the communicator finally verbalises what she intends to say. This last level also includes "the appropriate descriptions of entities and events" (Hobbs 1979, page 88). In this framework, entity status would be computed on the coherence

level, so that it is available to the sentence-level algorithm for generating referring expression. What is so interesting about Hobbs' approach is not this three-layer scheme, however, but the strong role that planning, and with planning, intentionality, plays in determining coherence.

3.2.2 Psycholinguistic Perspectives on Coherence

Hobbs' account focuses on producing coherent discourse. Most of psycholinguistics, on the other hand, has concentrated on how coherent discourse is processed (Schade et al. 1991). The basic heuristic here is: the more difficult a text is to process for an addressee, the more problems he has in establishing connections between the text's parts, and the less coherent the text.

What the basis for these connections might be is not quite clear (Harley 1995, Schade et al. 1991). Two approaches appear to be rather popular, Mental Models (Johnson-Laird 1983) and the Construction-Integration model of Kintsch (1988). In the Mental Models approach, the addressee continously updates a discourse model as new linguistic signs come in. This discourse model is a dynamic representation of the entities and events that are mentioned in the text. On the basis of their previous knowledge about these entities and events, addressees can infer information which is not explicitly present in the text. Cognitively, both explicit and inferred information is closely integrated, so that after a while, addressees have trouble distinguishing between inferences they have drawn and what they have actually read in a text (Hörmann 1979). This shows just how firmly addressees can integrate their interpretation of a text into their system of experiences, their personal experience theory, as Ungeheuer would call it (c.f. Appendix D). There are two ways of dealing with incoherence: Either a new discourse model is constructed, or the information is merely represented propositionally, without integrating it into the model.

In the approach of van Dijk and Kintsch (1983) (see also van Dijk 1980, Kintsch and van Dijk 1978), which was the basis for the more recent Construction/Integration model (Kintsch 1988), propositions are not integrated into models whose structure represent that of real-world events. Instead, Kintsch and van Dijk posit a far more abstract structure, which was developed by van Dijk (1972) for the analysis of texts in terms of a generative grammar. This structure consists of four levels of representation. Atomic propositions populate the first level. These are in turn hierarchically embedded into complex propositions, which form the second level. These complex propositions are linked on the level of *local coherence* in a graph. Taken together, all propositions that are contained in a text form the *text-base*. The propositions that are explicitly stated in the text are the *implicit* part of that text base. They are completed by additional propositions inferred from world knowledge to form the final *explicit* text base (explicit = explicated).

On the third level, the level of *macrostructure*, all propositions of a text are integrated into a coherent whole. Macrostructures represent how the meaning of a complete text is structured. Indeed, for van Dijk (1980, Section 2.3.1), texts are defined as only those sequences of sentences that have a macrostructure. Macrostructures need not be restricted to the semantic level; (van Dijk 1980, Section 3.4.12) explicitly extends the concept to structured sequences of speech acts. On the fourth and final level, we have *superstructures*. While (semantic) macrostructures serve to describe what a text is about, superstructures describe how that text is organised. In this way, they represent the highest level on which coherence can be established.

The current version of the Construction-Integration model implements a connectionist approach to coherence. Kintsch (1988) assumes that knowledge is not organised via fixed schemas or scripts, but as a connectionist network. The units of that network correspond to stored propositions. Incoming discourse is analyzed as follows:

1. represent the text by propositions. This translation is mandatory.

2. elaborate on the propositions. For each of the propositions, it is checked for which nodes from the general knowledge net it serves as a retrieval cue. This process is undirected.

3. generate additional inferences. In this step, macropropositions are retrieved that form part of the macrostructures of the original model; this is also where bridging is said to occur.

4. assign connection weights to the new connections from the new proposition units to the text base.

This is the *construction* phase. In the *integration* phase, these new units are then integrated into the knowledge network by repeated cycles of spreading activation and renormalisation along the connections in the net. If the net does not manage to stabilise, we go back to the construction phase and insert additional propositions.

Both Construction-Integration theory and Mental Models provide us with a view of both local and global coherence. When a text is locally coherent, the addressee can integrate incoming linguistic signs more quickly into the analogical discourse model / complex hierarchical structure that he uses to process the current discourse. While the Mental Model approach is computation oriented, giving a procedural account of how meaning is constructed, the structural approach provides an interesting bridge between semiotic and literary theories of text on the one hand, and psycholinguistic theories of processing on the other.

Both approaches emphasise that something like referential continuity is crucial for coherence. In the original (Kintsch and van Dijk 1978) paper that is the earliest commonly cited reference on the Construction-Integration model, referential coherence means that one and the same argument repeatedly appears in the propositions that are constructed during processing. Although they make it very clear that referential coherence is not the only way to establish a coherent text, it is the most important of the criteria they define. In that paper, they define referential coherence as mere argument overlap: two propositions cohere if they share at least one argument. For Johnson-Laird (1983, page 250), referring expressions are one of the three kinds of input for procedures that operate on mental discourse models, the other two are “the context as represented in the current mental model, and the background knowledge that is triggered by the sentence.”

Schade et al. (1991) have gone a step further and proposed a system-theoretic account of coherence. Both communicator and addressee belong to one large communicating system. If a text is to be coherent, both subsystems, that of the communicator and that of the addressee, must be in a stable state. The state of all subsystems involved in communication changes dynamically as a discourse is processed by the addressee and produced by the communicator. A text is said to be coherent if and only if it does not cause any of the subsystems that produce or process it to enter an instable, incoherent state. Coherence as a property of texts is termed “Objektkohärenz” (object coherence), coherence as a property of text reception and production is termed “Prozeßkohärenz” (process coherence). Applications of this concept can be found in

(Rickheit 1991). The main advantage of this view is that it relates coherence both to how texts are produced and to how texts are interpreted. The basic assumption behind this approach to coherence is that human communication can be modelled using the apparatus of mathematical systems theory and non-linear dynamics. In principle, there is nothing wrong with that; the approach has proved to be very useful in understanding—and in showing the limits of understanding — large-scale complex systems. However, I see both a mathematical and a conceptual problem here. The mathematical problem comes into play when the variables that enter into the model cannot be defined in terms of real numbers any more. The relation between discrete categories and continuous variables in non-linear dynamics does not appear to be well-understood (Leopold 1998). The conceptual problem I see is that the communication system is reduced to that of the communicator and that of the addressee. As far as I can see, the social systems of which both are part can only enter into the current model in the form of parameter settings. Although the system theoretical approach can be extended to social systems, as well, the resulting sets of equations would be monstruously complex. Therefore, even if we do manage to derive a set of mathematical equations for describing the processes of coherence, solving them might still be so difficult that we need to resort to simulations and shortcuts.

3.2.3 Summary

For the purposes of this thesis, I will therefore not adopt a systems-theoretic perspective. Rather, I will return to the model of communication described in Appendix D. If we accept that communication is an action, then the best way to describe how a discourse coheres is via the intentions of speaker and hearer. This high-level pragmatic approach, reminiscent of the solution of (Fritz 1982)

- covers coherence in written as well as in spoken language, as the speech act theoretic analyses of van Dijk (1980) and the results of Fritz (1982) show

- takes both the communicator's and the addressee's perspective into account. The communicator structures the sequence of signs that she will produce according to the intentions she pursues in communication, and the addressee interprets the signs that the communicator has produced on the basis of the working hypothesis that all signs contribute in some way or another to the intention behind the communicator's action, and that these signs can therefore be related to one another in some way.

- emphasises that coherence is a process. This is implicit in our dynamic model of communication. Each production and each interpretation of a sign is highly individual, a dynamic process that can never be repeated exactly, because it changes the individual system of experiences of each participant in subtle yet indelible ways.

To sum up, we have seen that classical text linguistic approaches to coherence and cohesion shows that coherence is itself multi-faceted, and that co-specification sequences are one of the means of establishing cohesion, surface indicators of coherence. Psycholinguistic research has shown that although co-specification is one cue to coherence among many, it is often very important for ease of processing. Correctly specified discourse entities which communicators keep coming back to during a discourse greatly help addressees to string more or less disparate

sets of utterances together. A more detailed look at how coherence is established shows that much depends on the communication situation and on the system of experiences of speakers and hearers. Structural entity status is thus a matter of how central speakers and hearers *judge an entity to be*, and a matter of the intentions with which they produce and process language. Entity status clearly contributes to cohesion because it influences how expressions that specify a discourse entity are coded linguistically. But it is not the universal key to coherence. Now that we have clarified this point, let us return to structural entity status proper: how can we describe the role that a given discourse entity plays in a discourse?

3.3 Discourse Structure

In this section, we will examine the relation between entity status and discourse structure. I have claimed in Chapter 2 that entity status describes in a theory-independent way what theories need to explicate. The following pages substantiate that claim: For three major theories of discourse structure, the theory of (van Dijk 1980) (Section 3.3.1), Rhetorical Structure Theory (Mann et al. 1992) (Section 3.3.2), and the theory of (Grosz and Sidner 1986) (Section 3.3.3), I show how to define structural entity status in terms of that theory.[2]

3.3.1 Van Dijk: Micro-, Macro-, Superstructure

We have already encountered van Dijk's theory in section 3.2.2 when we discussed the way it modelled coherence. It appears out of date to discuss such a venerable theory here, especially since modern computational linguistics largely ignores it. But modern psycholinguistics does not: In its 1983 application to discourse comprehension, (van Dijk and Kintsch 1983), it is one of the undisputed classics of the field, and, judging from more recent work by Walter Kintsch (1988, 1993, 1994, 1995), still very much alive.

In the structural framework of van Dijk, discourse entities can be conceived of as arguments of propositions, both atomic and complex. This explains the large ontological variety of discourse entities: the arguments of complex propositions can be propositions themselves, and they can—in principle—be arbitrarily complex. To define structural entity status in terms of van Dijk's theory is straightforward. Let me give the definition guided by the questions from Section 3.1:

1. *In which discourse segments does the entity occur?*
   The entity is realised, either implicitly or explicitly, every time it occurs as the argument of a proposition. If we assign each proposition a unique identifier, this property would reduce to the list of the identifiers of all propositions that the entity occurred in. When we want to determine how the segments in which the entity occurs are linked, we just need

[2] Since there are many competing approaches to discourse structure, it would be beyond the scope of this thesis to compare and evaluate them all. If I were to formalise structural entity status in terms of dynamic semantics, I would need to consider other models, such as the Questions under Discussion of (Ginzburg 1996) for dialogue, the rhetorical relations between discourse representation structures of (Asher 1993), or the tree-based Linguistic Discourse Model of Polanyi (1988).

to determine the place of the propositions in which the entity occurs in the macro- and in the superstructure.

2. *How is the entity linked to other entities in the discourse?*
   Entities are linked by the propositions they have occurred in together. This can be tracked by lists which state for each entity e_i that has co-occurred with an entity e_j how often the two have been arguments of the same proposition. If one assumes that the arguments of the propositions are sorted, one could also add information about the respective sort of e_j and e_i.

3. *How closely is the entity connected to the speaker's communicative intentions? Is it part of the gist of the discourse, of what the discourse is all about?* Since the theory does not explicitly model speaker intentions, except in pragmatic macrostructures, the first question is difficult to answer. The second, in contrast, is easy: since the semantic macrostructure represents the gist of the text, a discourse entity is part of the gist iff it is part of that macrostructure.

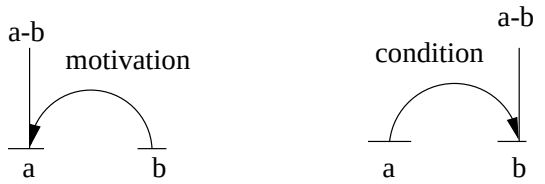
In practice, this analysis is not as easy as it seems. Although methods for text analysis have been developed that rely on Kintsch/van Dijk-style propositions (Früh 1998), analysing a text into its constituent propositions and determining the appropriate macro- and superstructure is an arduous task.

3.3.2 Rhetorical Structure Theory

RST models text structure by coherence relations between parts of a text. These parts are called *text spans*. After the analysis, the complete web of relations should provide a functional view of the hierarchical organisation of the text. The structure should cover all clauses of the text. At its top, there should be one main relation that covers the complete text. A detailed example for a RST analysis is provided in (Mann et al. 1992).

Rhetorical relations are described by schemata. These schemata list the text spans that are connected by the relation and the way in which they are related. All schemata are function-oriented, they describe "the work they [= text spans, M.W.] do in achieving the goals for which the text was written" (Mann and Thompson 1987, page 82). In other words, RST intends to describe how a writer has structured her text in order to achieve a communicative goal, how the information in a text is organised. Rhetorics and argumentation theory pursue similar aims (Ungeheuer 1974/1987b, Toulmin 1958), hence the terms *rhetorical* relations and *Rhetorical* Structure Theory. Typically, a schema involves two text spans: a nucleus and a satellite. A reader needs the nucleus in order to determine why the satellite occurs in the text, but not the other way around. Other types of schemata, which may also involve multiple nuclei, are discussed in (Mann and Thompson 1988, Mann et al. 1992). A sample analysis is presented in Figure 3.1, taken from (Moser and Moore 1996, Figure 3). For a more formal definition of the underlying relations, see (Mann et al. 1992). What the smallest unit of analysis should be depends largely on the analyst.

Fox (1987) investigated how rhetorical structure influences pronominalisation, concentrating on singular third-person references to persons. She found that several conspicuous patterns in her data, which consists of both multi-party conversations and written expository prose, can



Intentional Perspective Informational Perspective

Discourse:(a) Come home by five o'clock.(b) Then we can go to the store before it closes

Figure 3.1. Two RST Analyses of the Same Discourse. Source: (Moser and Moore 1996, Figure 3)

best be explained in terms of discourse structure. Incidentally, she used the fine-grained RST analysis only for written language; for speech, she applied conversation analysis, because this method is more suitable for the analysis of dialogues.[3] As of (Mann et al. 1992), the application of RST, which has been developed mainly on short written texts, to dialogue is still an open research problem. Although Fox' version of RST is based on an earlier version of RST, her results substantially still hold. Fox consistently uses propositions as units of analysis. For written expository prose, she found:

> A pronoun can be used to refer to a person if there is a previous mention of that person in a proposition that is active or controlling; otherwise a full NP is used.
>
> (Fox 1987, page 139)

An *active* proposition is a proposition whose partner in a rhetorical relation is currently being processed. This merely means that the antecedent of the pronoun occurs in the previous proposition, a result which confirms distance-based heuristics such as that of (Givón 1983a). Propositions are *controlling* as long as their partner in a rhetorical structure is active. In particular, this means that the proposition which contains the antecedent not necessarily needs to be adjacent to the proposition that contains the anaphoric pronoun. A special case of controlling propositions are created by *return pops*. Return pops close off an embedded discourse segment and return to a higher-level segment. Fox found that in written prose, pronouns were only used in return pops if the embedded segment was either very small or if the entity that the pronoun specifies was mentioned in the embedded segment, as well. A purely distance-based theory of pronominalisation can only account for the second case, but not for the first.

Fox' results show which aspects of rhetorical relations should go into a description of structural entity status. As in the Structure Theory of (van Dijk and Kintsch 1983), the structural status of an entity depends on the role of the propositions it occurs in. On this basis, we can now explicate structural entity status in terms of RST:

1. *In which discourse segments does the entity occur?*
   Locate the propositions where the entity occurs in the tree formed by the rhetorical relations. The segments in which the entity occurs are connected by rhetorical relations.

[3] How RST can be extended to speech, in particular dialogues, is still an open question.

2. *How is the entity linked to other entities in the discourse?*
   Explicit links are only specified between propositions. Without a more detailed model of the role of entities in propositions, such as conceptual structure (Jackendoff 1992), all we can state is which entities co-occur with which, both within a proposition and within a rhetorical relation.

3. *How closely is the entity connected to the speaker's communicative intentions?*
   Modelling the intentions of a speaker is a problem for RST (Moore and Pollack 1993, Moser and Moore 1996). RST relations can correspond to both informational and intentional functions. For example, the MOTIVATION relation in Figure 3.1 is intentional: the communicator wants to motivate the addressee to do something. On the other hand, sentence (a) is a condition for (b) to happen. The relation is clearly informational. Moore and Pollack (1993) argue that informational and intentional relations should not be conflated on the same level of structural description.[4] As a potential remedy, Moser and Moore (1996) suggest that the intentional structure of Grosz and Sidner (1986), discussed below, should be used for representing the intentional level, while RST should be confined to the informational one.
   Is it part of the gist of the discourse, of what the discourse is all about?
   This question is easy to answer. We just need to determine how the propositions in which the entity occurs are related to the highest-level rhetorical relations in the text. The idea is that the rhetorical relations which connect the largest text spans provide a brief summary of the main aim and content of the text. Marcu (1999) proposes a summarisation algorithm that exploits the nucleus-satellite relationships of RST.

Recently, Cristea and Ide (Cristea, Ide and Romary 1998, Ide and Cristea 2000, Cristea, Ide, Marcu and Tablan 2000) have proposed a new theory for deriving co-specification sequences from text, *Veins Theory*. The idea behind this theory is to restrict the search space for the antecedents of anaphoric expressions, be they pronouns or full NPs, on the basis of discourse structure, more precisely, RST. The domain which contains all permitted antecedents is called the *domain of accessibility*. This domain is described by *veins*. A formal definition of this notion is given in (Cristea et al. 1998). Informally, discourse entities that were mentioned in a nucleus remain accessible until that nucleus becomes part of a unit that functions as a satellite. The intuition behind this, already exploited by Marcu (1999), is that the information in nuclear units is more central than that in satellites. Nuclear units also have access to their left siblings. In the model of (Grosz and Sidner 1986), on the contrary, sibling discourse segments have distinct focus spaces which are not visible to each other. Therefore, an anaphora resolution algorithm that only looks for antecedents in available focus spaces would be at a loss when the antecedent of an anaphor is in such a sibling. Veins Theory allows to overcome that restriction (Ide and Cristea 2000). Veins Theory is essentially a formalisation of what I have called structural entity status in RST terms; it exploits the constraints that RST places on discourse structure (order of siblings in tree, distinction between nuclei and satellites) as far as possible. However, I doubt whether it is also an adequate model of how discourse entities are *managed*; for that purpose,

[4] Incidentally, because most RST relations are informational and not intentional, it has been suggested by systemic funcionalists that the theory be renamed *logical* structure theory (Hasan and Fries 1995a). In systemic functional theory, the level of semantics that informational relations are defined on is called the *logical* level.

cache mechanisms such as those proposed by Walker (2000) may also be necessary. Another problem with Veins Theory is that since RST allows multiple analyses of the same discourse, any implementation of the theory needs additional well-formedness constraints on the discourse trees produced (Marcu 1997).

Summary and Evaluation: In spite of its analytical power and extensive empirical testing, RST is by no means a perfect analysis tool. Two problems have already been mentioned in the preceding paragraphs: its limited applicability to dialogue, and the conflation of informational and intentional structure. But there are two even more fundamental problems with the relations and their definitions: Firstly, the names for the rhetorical relations are sometimes highly metaphorical, as the following quote shows:

> The elaboration relation is particularly versatile. An informal characterisation is:
> Elaboration: a satellite text span supplements the nuclear text span with one of the following kinds of detail:
>
> 1. set : member
> 2. abstraction : instance
> 3. whole : part
> 4. process : step
> 5. object : attribute
> 6. generalisation : specific
>
> (Mann and Thompson 1987, page 86)

No wonder that many researchers feel the need for more specific relations with more restricted definitions. Due to a tendency to tailor rhetorical relations to the needs of the system where they are used (mostly for text generation), the number of RST-type relation had already grown to more than 400 in 1993 (Maier and Hovy 1993). Mann and Thompson repeatedly stress that the fundamental set of relations is small (Mann and Thompson (1987) count 25 basic schemes), they state that the number of relations is in principle open-ended, and that their definitions are culture-specific. Sanders, Spooren and Noordman (1992) argue that this proliferation of coherence relations is psychologically unrealistic. They propose to model the set of relations that people recognise in texts by a limited number of features. Knott and Sanders (1998) found that the relations uncovered by this methodology largely overlap with relations that were found by a rigorous corpus analysis of connectives (Knott 1996).

The second main problem is that several rhethorical relations can apply to a stretch of text depending on how the annotator interprets it (Marcu 1997). Thus, it is perfectly possible that one annotator analyzes the discourse in Figure 3.1 using MOTIVATION, and another using CONDITION, as long as both annotators feel that the label is best suited to their reading of the text.

To sum up, although Fox (1987) has shown that patterns of rhetorical relations can explain some pronominalisation patterns quite well, there are enough problems with RST as it stands to cast doubts on whether RST is indeed an adequate approach to analyzing coherence relations. Moreover, as Fox' results on conversation show, the fine level of detail that RST offers may not

always be necessary. Sometimes, it is sufficient to have an idea of where discourse segments begin and end, and how they are nested. We will discover a theory that is mainly concerned with the relations between discourse segments in the following section.

3.3.3 Grosz & Sidner: Attention and Intention

Grosz and Sidner (1986) propose a theory of discourse structure with three levels: linguistic structure, intentional structure, and attentional state. Their starting question is: what makes a discourse hang together? In their paper, they integrate Grosz' work on task-oriented dialogues with Sidner's work on focusing to yield a theory that is intended to cover a wide range of types of discourse. With regard to coverage, Grosz and Sidner are more ambitious than RST, which started small with short edited texts (Mann and Thompson 1987), but when it comes to defining structures and relations between them, their framework is much more sparse.

The linguistic structure is the basis. It consists of utterances, which in turn are structured into discourse segments. Grosz and Sidner (1986) assume that discourse segments can be determined reliably: subjects that segment the same discourse will place the boundaries between roughly the same utterance. Passonneau and Litman (1993) confirm that intuition, although rough is indeed the correct term for the level of agreement that they found. Many boundaries in their data were only marked by one or two subjects, but there were also a number of boundaries on which most subjects agreed. Overall, the level of agreement on a boundary tended to be either very high (most of the subjects had detected it), or very low (only one or two subjects saw a boundary there). This shows that some boundaries and hence some segments are more easy to recognise than others.

But what makes a string of utterances coherent? Grosz and Sidner's answer is: Because the communicator pursues a — preferably single—intention with that segment. They argue that only intentions provide a sufficiently rich basis for explaining why a discourse is structured as it is. It is not possible to describe discourse with a fixed number of semantic relations. Each discourse segment (DS) is characterised by one main intention, the *discourse segment purpose* (DSP). To determine a DSP is to specify what exactly is being intended by whom. The DSPs are connected by two structural relations, dominance and satisfaction-precedence. A discourse segment purpose DSP1 of a discourse segment DS1 *dominates* the discourse segment purpose DSP2 of another segment DS2, if DSP2 contributes to achieving DSP1. Satisfaction-precedence is motivated by the analysis of task-oriented dialogue. If DSP1 must be achieved (or, in Grosz and Sidner's terms, satisfied) before DSP2, then DSP1 *satisfaction-precedes* DSP2.

The attentional state is the component of the model which has perhaps received the most attention in the literature, not least because it is the basis of Centering Theory (Grosz et al. 1995). The attentional state models the participants' focus of attention. It is dynamic; it keeps changing as the discourse progresses. Attentional state is modelled by a sequence of *focus spaces*, which, together, form the *focusing structure*. A focus space contains all properties, objects, and relations in a discourse segment. To put it the other way around, each discourse segment has its own focus space, which is continuously filled while the segment is produced or processed. After a discourse segment has been finished, its focus space is removed from the focusing structure. That structure is realised as a stack: last in, first out. Accordingly, the operation that removes a focus space from the stack after it has been completed is called *pop*, and the operation that creates a new focus space on the stack as a segment is opened is called

push. If a new segment S_2 is opened within another discourse segment S_1, then the focus space of S_1 is not popped from the focus stack. Instead, the focus space of S_2 is popped onto that of S_1. When the focus space of S_2 is popped off, because S_2 has been closed, the space of S_1 moves to the top again. This is called a *return pop*.

Attentional structure is the level at which entity status comes into play. Let us reconsider our three questions again:

1. *In which discourse segments does the entity occur? How often? Where exactly?*
   An entity occurs in a discourse segment if it appears in the focus space of that segment. Its status within that segment changes dynamically; these local changes are tracked by the center list. Roughly, a center corresponds to a discourse entity that a pronoun in the current utterance can specify.

2. *How is the entity linked to other entities in the discourse?*
   The GS theory provides data about the segments in which the entities co-occur, and how often they have been in the focus space together. Other factors such as the thematic role or the grammatical function of the expressions that specify the entities are also available, because they are needed to compute the order of the list of forward-looking centers for Centering, the associated theory of local coherence.

3. *How closely is the entity connected to the speaker's communicative intentions?*
   Since the whole description of discourse structure that GS provide is intention-based, the answer is straightforward: analyze the DSPs of the segments in which the entity occurs and their relations both to each other and to DP, the purpose of the complete discourse.
   Is it part of the gist of the discourse, of what the discourse is all about?
   Grosz and Sidner offer a very interesting take on the old notion of "topic":

   > It appears that many of the descriptions of sentence topic correspond (though not always) to centers, while discourse topic corresponds to the DSP of a segment or of the discourse. (Grosz and Sidner 1986, page 192)

   According to this quote, discourse entities can only be sentence topics. However, if they occur in the DS whose DSP represents the discourse topic, and if they are topical in that DS, they really are important for describing the gist of the discourse. As we will see below (Section 3.4), this distinction makes more sense than frantically looking for NP discourse topics where there are none.

The GS model has been very influential. In the literature on referring expressions, it is almost always merely introduced as the backdrop of Centering Theory, which is discussed in more detail in Section 4.3.2. But Centering is just concerned with local coherence inside a discourse segment. The aspect of global coherence has been rather neglected. One of the rare exceptions are (Hitzeman and Poesio 1998). They examine how the focus stack (not just the current focus space) can be used to resolve long-distance pronouns. Walker (1998) proposes to replace the hierarchical focus stack structure with a *cache* which stores the last seven or so discourse entities. She motivates this additional data structure by analyses of return pops and

pronominal coreference across discourse segment boundaries. Although she may be correct in that the additional machinery of discourse segments and their focus stack is not needed in order to explain pronominalisation, this argument is not sufficient to discard the whole theory of discourse segments and focus spaces. After all, there must be other test beds for theories of discourse structure than the well-worn staple of pronoun resolution.

3.3.4 Summary

Undoubtedly, Grosz and Sidner have developed a very powerful theory, albeit difficult to implement and annotate. This is not surprising: recognizing intentions in discourse is hard. However, as Section 3.2 has shown, intentional structure is extremely important for establishing and maintaining coherence. Therefore, as a formal approach to global coherence, the GS model appears to be on the right track. The discussion of RST in Section 3.3.2 has shown that Grosz and Sidner's reserves against a fixed set of discourse relations and indeed against any purely semantically motivated relations are justified. Still, for the purposes of text analysis text generation, a more detailed taxonomy of coherence relations is indispensable.

RST provides a tried and tested framework for analyzing discourse in terms of coherence relations. However, mere informational coherence relations are not sufficient; we do need some information about intentional structure if we are to analyse more than short prose texts. A crucial problem with RST is that the relations are highly metaphorical, and therefore many definitions are just not clear. A good theory should be easy to extend to discourse, and it should also be easy to relate to the schemata that many analysts have found useful in cognitive science (Schank 1977). As to the number of relations, I doubt whether psycholinguistic evidence can uncover a fixed, small set of relations which govern the way humans interpret coherent text. From a communication theoretic point of view, it appears highly plausible that connections which establish coherence are made up and adapted on the fly. Although there are common schemata, which are slightly modified each time they are used to process texts, people are flexible enough to create new relations on the fly if these schemata should fail them.

Structural entity status needed to be explicated slightly differently for each of the theories. Since both Fox-style RST and van Dijk assume that the smallest unit is the proposition, it follows that structural entity status must be computed from the place of an entity in the hierarchy of propositions defined by the discourse structure. Grosz and Sidner, on the other hand, concentrate on discourse segments. Structural entity status is largely taken care of by a special data structure, the focus stack. The position of an entity on this stack and its relative salience are the main parameters that need to be stored. But in fact, what we need is a merger of those two kinds of information. We need to know both the role of the proposition (or the utterance) in the discourse and the position of the entity on the focus stack. We get both kinds of information in the merged discourse theory that Moser and Moore (1996) describe. Ultimately, an integration of coherence relations with the intentional model of Grosz and Sidner although this might lead to a much more concise notion of structural entity status.

Finally, it may well be that even the best theory of discourse structure cannot explain the form of certain references to entities which have last been mentioned dozens of pages ago. Let me cite a prototypical example:

> How long does definiteness last? On page 13 of Arthur Koestler's *The case of the midwife toad* there is the sentence *One of his [Paul Kammerer's] pockets contained a letter addressed 'to the person who finds my body'*. The letter is not mentioned again until page 118, where we find a sentence beginning *As far as we know, he wrote four farewell letters (apart from the note 'to the person who will find my body')* ... Thus, with the aid of the quote from the letter, the reader is assumed to be able to identify what letter this is, and we can say that the status of definite for this referent has been preserved over 105 pages.
>
> (Chafe 1976, page 40; underlined words in original replaced by italics)

Chafe goes on to speculate about the connection between the two passages: the two of them frame a flashback to the "events which led to Kammerer's suicide" (op. cit.). We will return to such examples in chapter 4 on the management of discourse entities. But until we can turn to the management dimension, we need to discuss one more well-known concept of analysis that can be related to structural entity status: the notions of "theme" and "topic".

3.4 Theme and Topic

Of the three questions about properties of structural entity status that I asked at the beginning of this chapter, the first two are clearly related to discourse structure. The last one, however, points in the direction of another popular concept in linguistics, the notion of *theme* or *topic*. The literature is replete with different definitions of these terms. Theme and topic have been used as analytical tools by researchers from all linguistic traditions and specialisations. Therefore, any overview of uses must be biased by the traditions that the authors of that overview are familiar with. My bias is general and computational linguistics, and even in that field I do not strive for a complete overview. Instead, I aim to show how something like entity status might be integrated into some popular directions of research.

This section is structured in a similar way as Section 3.3. Instead of developing an all-new improved notion of theme for the new millenium that incidentally corresponds to what I have called structural entity status, I will survey several definitions of topic/theme and relate them to structural entity status. The discussion will be guided by the questions listed on page 24: *How closely is the entity connected to the speaker's communicative intentions? Is it part of the gist of the discourse, of what the discourse is all about?* Clearly, structural entity status is influenced by the discourse as a whole, so, ultimately, it will have more to do with discourse theme. But this discourse-level notion will need to be related to sentence-level themes, as well. Therefore, in section 3.4.1, I will review approaches to sentence theme. If we see themes as links to the co-text, the connection between theme and structural entity status is clear: entity status protocols how often an entity has served as link, and when. Usually, the dichotomy between psychological subject and predicate is also discussed in the context of sentence theme. But this procedure does not do the scholars who developed these notions justice. To argue this point in more detail would lead us too far afield here; details can be found in (Wolters in preparation). Section 3.4.2 links the discussion of theme to Section 3.3—it is about discourse topics and how these topics relate to structural entity status, which depends of course greatly on how discourse structure is modelled. Finally, in Section 3.4.3, I summarise how discourse topic and sentence theme can be related to structural entity status.

A Note on Terminology: When discussing the research of others, I will always use the term that is favoured by the researchers I am citing. When summarizing different research traditions, I will use "theme" when talking about sentence-level theme or topic, and "topic" when talking about the discourse level. For the sentence-level, I favour "theme" for two reasons: first, its counterpart, "rheme", is less polysemous than one of the two counterparts of "topic", "focus", second, it is mainly used by researchers who see themes as contextual links. For the discourse-level, I will use the term "topic". This is the commonly used term in English language publications (Brown and Yule 1983). Common English usage is the only reason why I chose "topic" over "theme" here—I would not hesitate to translate my "discourse topic" into German as "Diskursthema".

3.4.1 Sentence Theme

Almost all recent papers on theme distinguish a sentence-level version from a discourse-level version.[5] The sentence-level theme is frequently defined from a pragmatic point of view. It is something "given", "old", which the sentence is "about" or which can be taken to be the point of departure for interpreting the sentence. Sentences not necessarily have to have themes. They can also consist of all-new information, showing no direct link to the previous discourse context. In sentences with a theme, the communicator arranges the information so that the rheme is predicated of a theme. Sentences without a theme merely present a state of affairs. The first kind of sentences are commonly called *categorical*, the second kind *thetic* (for more on this distinction, c.f. Sasse 1987).

Some researchers have argued that both terms, topic and theme, can and should coexist on the sentence level, so that they can share the workload of meanings which has been attached commonly to just one of the pair. For example, Halliday explicitly characterises only one part of his theme, that part which has an experiential metafunction, as truly topical. Molnár (1993) goes a step further. Based on Bühler's (1934) Organon model, she defines three relevant dichotomies:

Topic-Comment:	topic is what comment is about.	(pragmatic aboutness, *Darstellungsebene*)
Theme-Rheme:	theme known to the hearer, rheme not	(givenness, *Appellebene*)
Focus-Background:	focus is what speaker judges to be relevant	(relevance, *Ausdrucksebene*)

The focus-background dichotomy is very popular in formal semantics (prominent recent examples are von Heusinger 1999, Dekker 1998, Krifka 1992, Büring 1996). The classical reference on this dichotomy is (Jackendoff 1972). Researchers have used it mainly for describing the meaning of intonational focus: The focus is the part of the sentence that is in the scope of a focus operator, while the rest of the sentence constitutes the background. I will not explore the relation of focus/background to entity status here, because that would lead too far afield.

[5] The distinction is sometimes attributed to (Reinhart 1981), sometimes to (van Dijk 1977), but the basic insight that sentence themes are not what a complete discourse is about is much older, and it would be foolhardy to try and trace it to one single pioneer—not least because this insight is obvious. However, it is certainly correct to say that Reinhart and van Dijk *popularised* the distinction among generative grammarians and formal semanticists.

Instead, I prefer to concentrate on more traditional approaches to information structure, which have been developed from metaphors such as contextual boundedness or point of departure. For analyzing discourse, we need many perspectives; focus/background is but one of them. In the following example, Rincewind is both intonationally focused and functions as a contextual link:

(3.5) Every child on Discworld knows that Rincewind is a shoddy magician with a strong instinct for self-preservation that some would call cowardice.
Why on earth did the City Council then choose [RINCEWIND]$_F$, of all people, for a reconnaissance mission to the Counterweight Continent?
When a messenger came to bring him the good news, Rincewind was shocked. His first thought was to flee the city.

What is the topic of the second sentence? The only serious candidate for that job is the candidate for the reconnaissance mission, Rincewind, who is incidentally also in the focus of both intontation and the City Council of Ankh Morpork.

Linguistic correlates of sentence theme such as word order, syntactic topic markers, most notoriously Japanese "wa" (Kuno 1972), and special topicalisation constructions (Lambrecht 1994) have been discussed extensively in the literature. For these sentence-level studies, researchers needed a notion that helped them explain what the constituents that appeared in that place in the sentence, had been dislocated, or with a certain particle, had in common, and how their common property could be related to communication. Based on their interpretation of the language samples they studied, frequently isolated sentences, they found paraphrases of the term "theme" that allowed them to capture that common property, and to formalise syntactic properties of topical constituents. For example Molnár (1993) models syntactic topics in Hungarian as adjuncts of focus phrases. Since she defines topics in terms of pragmatic aboutness, topic-comment structure is therefore right at the interface between the syntactic and pragmatic modules of a generative grammar (Motsch, Reis and Rosengren 1990).

Another motivation for the notion of sentence-level theme is theoretical. Scholars such as Paul (1920) or Wegener (1885), pondering how and why people can understand texts, found an answer in the dichotomy between *psychological subject* and *psychological predicate*, where the psychological subject serves as some sort of basis for interpreting the psychological predicate. The two motivations also differ in the role which they assign to notions such as "theme". Those who are mainly interested in powerful conceptual tools for describing certain aspects of language will be inclined either to jettison the term completely, because it is not sharp and concise enough (e.g. Schlobinski and Schütze-Coburn 1992) or to dissolve the term into a set of features which describe the functions that have been attributed to them (e.g. Jacobs 1999, see especially his typology of research attitudes to the topic-comment distinction). On the other hand, those researchers who believe that there is a stratum of linguistic systems which is called information structure, and that this level of structure specifies how the content of messages (= the information these messages carry) is adapted to the current discourse context, will face the chaos and modify the terminology to suit their needs and the linguistic theory that forms their background. Of course, both currents cannot be separated as neatly as I have suggested. The frequent citations of (Paul 1920) show that most linguists who have worked more extensively on theme are aware of his concepts, and researchers may believe in something like information structure even if they refuse to talk about it in terms of topics, themes, comments, foci, or backgrounds.

The vicious circle of conceptual polysemy was set in motion when researchers started citing each other's concepts of "theme" with little regard of the paraphrases and metaphors behind these concepts. Schlobinski and Schütze-Coburn (1992) and Vallduvi and Engdahl (1996) provide good overviews of the conflicting and complementary views and traditions, and Hasan and Fries (1995a) show how Halliday's concept of theme has been misinterpreted by those researchers who neglected its functional underpinnings.[6]

In this section, I cannot strive for a complete coverage. Instead, I will focus on four approaches to "theme": point of departure, aboutness, contextual links, and communicative dynamism. As in the section on discourse structure, I will leave formal semantic approaches to theme and its mother notion, information structure, aside. For important recent work, see e.g. (Büring 1996, Jacobs 1999, Roberts 1997, Steedman 2000a).

Aboutness: In recent years, Reinhart (1981) has undoubtedly been the most influential advocate of defining themes as that which a sentence can be interpreted to be about. Of all the approaches we will examine in this section, the aboutness approach has the most direct link to entity status, because sentences are typically analysed as being about *discourse entities* (Davison 1984, Lambrecht 1994, Gundel 1985, Gundel 1988). Most researchers in the "aboutness" tradition use the terms topic/comment to express the partition they are describing. A major motivation for this strand of research are the "topicalisation" constructions, such as left-dislocation of constituents in English. A typical definition is the following:

Definition 3.1 (Topic) *An entity, E, is the topic of a sentence, S, iff in using S the speaker intends to increase the addressee's knowledge about, request information about, or otherwise get the addressee to act with respect to E. (Gundel 1988, page 210)*

Definition 3.2 (Comment) *A predication, P, is the comment of a sentence, S, iff, in using S the speaker intends P to be assessed relative to the topic of S. (Gundel 1988, page 210)*

The definition of topic makes the notion of "aboutness" more explicit. It also implies that topics are particularly important for processing sentences (Davison 1984). The definition of comment, on the other hand, alludes to the scene-setting function of topic. Both definitions focus on language as a medium for exchanging information between rational agents.

Aboutness is not to be confused with givenness. For a very detailed discussion of this point, see (Lambrecht 1994). There is overlap: Topics are usually familiar to both speaker and hearer, and hearers can usually uniquely identify the discourse entities that are topical (Gundel 1988). But the two dimensions are not parallel, because familiar entities can be part of the comment. The classic example is based on reflexive anaphora:

(3.6) Who did Rincewind hurt when he cast the spell?
He hurt [himself]$_{\text{Comment}}$.

[6] Incidentally, Vallduví, who used to advocate a tripartition of information structure into link, tail and focus, where link and tail correspond to the background, is now mainly working with the notions of "rheme" and "kontrast" (Vallduví and Vilkuna 1998).

In terms of entity status, topicality belongs clearly to the structural dimension. If we track how often an entity has belonged to the topic, and if we add information about the distance of the sentences in which the entity has occurred as topic as well as information about the syntactic construction by which that entity is realised, then, we get a very detailed picture about the role that the entity plays in the informational structure of the discourse. Givenness, on the other hand, belongs firmly to the management dimension: it characterises how quickly the entity can be retrieved, if at all.

Sentence topics are based on a relation between propositions and discourse entities:

> A referent is interpreted as the topic of a proposition IN A GIVEN DISCOURSE the proposition is construed as being ABOUT this referent, i.e. as expressing information which is RELEVANT TO and which increases the addressee's KNOWLEDGE OF this referent.
>
> (Lambrecht 1994, page 127)

The important difference between the two superficially similar definitions of Lambrecht and Gundel lies in the three words "is construed as". For Lambrecht, the easier a discourse entity is to access mentally, the easier it is to interpret a proposition as being about that entity. Lambrecht's notion of accessibility will be discussed further in Section 4.3.1.

Lambrecht's definition contains another important insight: topics are construed by hearers when they interpret sentences. Let us take this statement one step further: If the topic is not marked either syntactically or morphologically, the hearer is—in principle—free to interpret the proposition to be about whatever he chooses, provided that this interpretation does not conflict with the context of the discourse. Unsurprisingly, it turns out to be rather difficult to apply the aboutness criterion to naturally occurring discourse, when the question test is not available anymore. To remedy this problem, formal semanticists such as (van Kuppevelt 1996, Ginzburg 1996) have suggested to represent discourses in form of question-answer pairs. Klein and von Stutterheim (1992) and Lötscher (1987) have shown that the question-answer paradigm can even be extended to determine a discourse topic. For them, the topic of a discourse is the desire for information to which it is supposed to satisfy, the kind of information that addressees hope to get out of the text. I am rather skeptical about such tests. Although it may be possible to artificially construct questions for task-oriented dialogue or expository text, an approach such as that of Grosz and Sidner, where intentional structure plays a central role, is both more general and more fruitful.

Komegata (1999) points out that the problem with the question test is only a theory-specific instance of a more general problem, the *identification problem:* How can we determine the topic of an arbitrary sentence on the basis of a given definition of topic? From the communication theoretic perspective described in Appendix D, we need to answer this negatively: it is not possible to give a such a general procedure, because each instance of use differs subtly from all others. Such a position can be derived plausibly from the nineteenth century concepts of *psychological subject / psychological object*, which we cannot go into further here (c.f. Wolters in preparation).

Point of Departure: With each sentence they utter, speakers follow a specific goal with respect to the discourse. In their sentences, they relate this goal to a point of departure, which is

shared by both speaker and hearer. This interpretation of the underlying structure of sentences, proposed by Weil (1844/1978), provides the literal point of departure for many current theories of theme and rheme. This metaphor is also central to the distinction between theme and rheme that Halliday makes in his seminal (1967) paper. He defines the theme as a category that structures "the clause as a message". Halliday's notion of theme is influenced both by the "psychological subjects" of nineteenth century grammarians and by Prague School work such as (Mathesius 1929). Halliday defines psychological subjects as follows:

> Psychological Subject meant 'that which is the concern of the message'. It was called 'psychological' because it was what the speaker had in his mind to start with, when embarking on the production of the clause. (Halliday 1994, page 31)

The second sentence of this quote contains in nuce the "point of departure" of the following definition of theme:

> The Theme functions in the structure of the CLAUSE AS A MESSAGE. A clause has meaning as a message, a quantum of information; the Theme is the point of departure for the message. It is the element the speaker selects for 'grounding' what he is going to say. (Halliday 1994, page 34)

When we compare the two quotes, we already see the vicious circle at work. What the speaker has in mind when she starts to produce a clause need not be the theme; this definition could also apply to the goal the speaker pursues with her utterance, the point she intends to make. In contrast, Halliday's definition of theme refers to how the message is structured. The point that the speaker wants to make is underlined by the thematic structure she chooses, which constituents she uses to set the scene for what comes next.

In English, this scene-setting function is accomplished by word order. The theme of a clause consist of the constituent(s) which come(s) first in the clause. Many researchers have criticised that Halliday chose to associate a functional definition with a fixed position in the clause. But if one looks more closely at the metaphor which he chose for paraphrasing the function of themes, this association makes perfect sense. What comes first in a clause is that which is processed first, and indeed, it sets the scene for the processing of what is to follow. Figure 3.2 illustrates how complex themes can be. Single themes belong to the experiential level. Unmarked themes correspond to standard word order, marked themes to non-standard word order, in particular, left displacements.

The exact definition of Theme in SFL is hotly debated. For a summary of the debate, see (Hasan and Fries 1995a), and for attempts to link the notion of theme to text structure, (Fries 1995, Ravelli 1995). Since the definition of theme is a metaphor, and since this metaphor has been rephrased frequently in SFL, the most urgent task according to (Hasan and Fries 1995a) is to reach a broader consensus about what the cross-linguistic function of theme is. I find the metaphor of "scene setting" easier to apply than that of "aboutness". Moreover, the concept can be linked directly to psychological theories of incremental processing. What comes first influences how the rest is interpreted; this intuition is covered superbly by the way theme is realised in English. Finally, Halliday's definition of theme is not restricted to content, he also

Metafunction	Definition	clause as . . .	components of theme
textual	creating relevance to context	message	continuative structural (conjunction or WH-relative) conjunctive
interpersonal	enacting social relationships	exchange	vocative modal (adjunct) finite (operator) WH- (interrogative)
experiential	constructing a model of experience	representation	topical (participant, circumstance, process)

Example:

well	but	then	Ann	surely	wouldn't	the best idea	be to join the group
continuative	structural	conjunctive	vocative	modal	finite	topical	
textual			interpersonal			experiential	
Theme							Rheme

Figure 3.2. The structure of a theme with multiple parts, based on (Halliday 1994, Table 2(4), page 36, Table 3(7), page 54, and Figure 3-13, page 55). Metafunctions are levels on which language can be described functionally. The *textual* metafunction describes how a text is organised. The *interpersonal* metafunction is concerned with how the interaction between communicator and addressee is organised. Finally, the *experiential* metafunction focuses on how language is used to structure and describe reality.

takes into account how the current sentence fits into the text as a whole (textual level) and into the current interaction between communicator and addressee.

Discourse entities can only occur as topical themes. The structural entity status variable records how often an entity occurs as a topical theme in the texts as well as the distance between these occurrences. In SFL, topical themes are used to chart the thematic progression of a text (c.f. also the sample analysis in Halliday 1994). However, it is not possible to deduce from the thematic progression the points that the communicator wishes to make; it merely offers a skeleton of how these points are strung together, and structural entity status describes the role of discourse entities in that skeleton. For example, in Halliday's analysis, the isolated sequence of themes make it clear that in the first two paragraphs, Robert is the main protagonist, while the third to fifth are organised around George, Robert's father. It is not clear to what extent this effect is preserved in languages with zero anaphora, such as Chinese and Korean, or in languages where the surface word order is not SVO, such as Welsh or Gaelic. In such languages, thematic progression chains would need to be analyzed from a different perspective. Although we can circumvent these problems to a certain extent if we asssume that structural entity status is mainly determined by the cohesive relation of Reference as defined in (Halliday and Hasan 1976), the degree to which an entity can serve as "point of departure" is nevertheless an important parameter which any account of structural entity status should cover.

To sum up, although Halliday's approach is intuitively appealing, the definition of theme is too vague to be a really useful tool for analysis, and most of the examples I have seen so far come from SVO languages with (almost no) zero anaphora. Hasan and Fries (1995a) are

remarkably frank about the reason: Concepts for linguistic analysis are always also described with respect to at least a few formal properties, and these properties can differ slightly from language to language. Hence, each language has to be investigated on its own terms, and each researcher needs to evaluate critically how the abstract, highly metaphorical descriptions that SF theory offers should be applied to her data.

Link to the Context: One of the oldest approaches to theme and rheme is based on the metaphor of "contextual link" or "contextual boundedness". Before something can be a contextual link, it has to be given or known by the context. For this reason, many definitions of theme are phrased in terms of givenness - such as the definition given by Ammann (1928), who is commonly credited with inventing the theme/rheme dichotomy. For Ammann, the theme is that which is already known, while the rheme is that which is new, the information which the communicator can give to the addressee about the theme.

Nowadays, the view of themes as contextual links is closely associated to the Prague School of functionalism. Contrary to what most brief discussions of the Prague School suggest, this research tradition is neither monolithic nor does it pursue a radical functionalism as caricatured in (Givón 1995b). It is firmly anchored in traditional structural analysis. Sgall, one of the main protagonists of the Modern Prague School, puts it this way:

> One of the substantial aspects of the functional approach thus consists in identifying the empirically established units of the system of language, oppositions present in the language as a system. These units and oppositions are established on the basis of operational (testable) criteria, i.e. it has to be shown that each of the postulated oppositions plays a role (has a function) in the relevant position of the patterning of language. Only such units can be established as are really needed, and whose presence is useful for the description as a whole.
>
> (Sgall 1987, page 169f.)

The classic approach that researchers tend to cite when discussing the Prague School is the Functional Sentence Perspective (FSP) (Firbas 1974, Firbas 1992). FSP was first discussed by Mathesius (1929), whose work was in turn inspired by (Weil 1844/1978).

The original aim of FSP analysis is to show how the flow of ideas in a discourse is expressed by grammar. In (Firbas 1974), FSP is one of three perspectives under which a sentence can be described, the other two being semantic and grammatical structure. In Mathesius' work, the theme can be both that which is spoken about and what is already known. Firbas redefines theme and rheme on the basis of *Communicative Dynamism* (CD). CD is a gradient notion. The degree of CD carried by a constituent is "the extent to which [it] contributes towards the development of the communication" (Firbas 1974, page 19). The theme proper is the constituent with the lowest degree of CD, the rheme proper has the highest. The constituents between theme and rheme form the *transition*. CD is reflected in word order, but only as far as language-specific grammatical rules permit. For example, in Czech, a language with a relatively free word order, the surface word order mirrors CD quite well. In English, on the other hand, surface word order and CD correspond far less closely because of the rigid SVO scheme (Mathesius 1929).

There is a close relation between CD and contextual boundness. This relation is exploited heavily by the Modern Prague School. Although CD still plays a role in their take on information structure, Sgall, Hajičová and Panevová (1986) prefer to describe information structure

with respect to Contextual Boundedness. In the Functional Generative Description (FGD) approach of (Sgall et al. 1986), information structure is described by the *Topic-Focus Articulation* (TFA).[7] TFA is defined in terms of tectogrammatical representations (TR), a dependency-based description of the structure of a sentence. Discourse entities that are realised by referring expressions correspond to an argument node in TR. Communicative dynamism describes the linear order of the argument nodes. This order can differ from surface order if the sentence accent is not placed on the rightmost constituent, or if shallow movement rules have moved an element to a different position in order to satisfy language-specific constraints on word order. This definition of CD mirrors the operationalisation that Firbas originally defined in Weil's terms. For each set of arguments, there is an unmarked *systemic ordering* (SO), which is defined in terms of thematic roles. A constituent is contextually bound if it comes earlier in CD than it would in SO. Again, the topic proper is the least communicatively dynamic, and the focus the most dynamic item. Since the definition of TFA relies heavily on a full tectogrammatical parse, empirical analyses of TFA in naturally-occurring discourse are rather time-consuming.

Like aboutness, both TFA and the more traditional theme-transition-rheme partition can be related easily to structural entity status, because both partitions are defined on syntactic representations. This permits us to derive the topicality of a discourse entity directly from the syntactic properties of the corresponding referring expression. Contextually bound discourse entities provide links to the preceding discourse. To make the link between sentence-level theme and structural entity status complete, we need some kind of bookkeeping which tracks when which entities are thematic.

Such a bookkeeping is made much easier by an adequate typology of sentence transitions like that developed by (Daneš 1974a). He analyzes the sequence of themes in the sentences of a discourse in terms of *thematic progression* of a text. There are three main types of transitions between two sentences S_1, S_2 with themes T_1, T_2 and rhemes R_1, R_2:

1. simple linear thematic progression: $T_2 = R_1$.

2. thematic progression with a constant theme: $T_1 = T_2$

3. thematic progression with derived themes: $T_2 \neq T_1$ and $T_2 \neq R_1$. In this case, subsequent themes often relate to parts of R_2.

Apart from these transitions, two more constructs are needed: jumps, which result when themes are omitted, and inserted material, which distorts the original form. Daneš' approach is systematic, but as Dressler (1974) remarks, it is only the first step to incorporating FSP in discourse structure, and at the time of writing, it appears that most of the remaining steps still need to be done. Researchers tend to use Daneš' categories more as a substitute for a more elaborate discourse structure, than as a link between sentence structure and text structure.

The "contextual link" metaphor has also been used in the context of formal pragmatics (Vallduvi 1990) and computational linguistics (Komegata 1999). In his thesis, Komegata presents a fully formalised concept of *conceptual link*. This formalisation is based on Combinatory Categorical Grammar (CCG). Steedman (2000b) has shown that CCG is powerful enough for an integrated account of the interaction between syntax, semantics, and information structure.

[7] The following summary is largely based on (Sgall et al. 1986, Kruijff-Korbayová 1998).

Komegata's approach has been implemented in a system that analyzes the information structure of English sentences and translates them into Japanese, which has overt markers of information structure, the particles *-wa* and *-ga*. Komegata defines theme and rheme on the level of propositions. The partition is strictly binary; propositions result from the semantic composition of a theme and a rheme. Themes are necessarily contextually linked, but need not be contrastive, while rhemes are necessarily contrastive, but need not be contextually linked. Contextual links are defined on the level of discourse entities. They relate entities in the current utterance to entities that are either present in the communication situation or have already been mentioned in the text. The inferences that addressees need to draw in order to establish that link are bounded, but these bounds depend on various factors which are not part of the inference logic itself: linguistic marking, accessibility (the management dimension of entity status), and domain knowledge. Like Lambrecht, Komegata thus ties thematicity closely to givenness. Since he restricts his link definition to discourse entities, structural entity status can be derived as usual: protocol how often an entity acts as the anchor point for a contextual link, and which relation exists between theme and anchor.

Summary: We have seen that modern approaches to sentence theme vary greatly. Most researchers, especially the formalists, prefer to define only referring expressions as themes. But as we will see below and as we have seen in Section 3.3, this restriction quickly runs into problems when the information structure of a sentence has to be tied to the informational or intentional structure of a discourse, because both are organised in terms of propositions, not in terms of discourse entities. What is a sentence about? What is a discourse about? Clearly, both questions have to be answered on the level of semantics. How do we determine what a sentence or a discourse is about? These are questions for pragmatics—how is language used to mark those constituents that play a special role in structuring the content of a discourse? I think that one question cannot be answered without the other, and that sentence theme and discourse topic cannot be defined independently. Once we know the position and the function of a sentence in its context, we can determine how it is linked to what precedes and what follows it, how it is embedded in the current communication situtation, what it is supposed to be about. If this link is missing, we can either follow the inductive strategy of (Jacobs 1999), who splits the topic/comment distinction into a set of four features which describe the contexts in which certain syntactic constructions can appear, or we can take our refuge in those parts of a sentence which can be linked to context more easily than propositions: discourse entities. The first strategy certainly has its merits, if the aim is to develop inductive, well-founded categories for syntactic analysis. But the original functional motivation for something like themes is lost. The second strategy is also promising, and can be formalised reasonably well. But it opens the doors wide to people who are tempted to confuse givenness with thematicity.

I find the idea of topics as contextual links very promising. It answers several questions in an elegant way:

1. Does every sentence have to have an explicit topic?
   No. If the hearer can infer the necessary connections from the context, there is no need.

2. Do topics have to be referring expressions?
   No, but they should be discourse entities, because topics are anchor points, and therefore hearers should be able to refer back to them.

3. Does it make sense to assume degrees of thematicity?
   Yes. It does not make sense that only a certain part of a sentence links that sentence to the context. All constituents can serve as links. The more difficult it is to find an anchor point for a contextual link, the less thematic is that link. The theme proper of a sentence is the constituent with the clearest link.

From this short overview, it becomes clear how theme can be integrated into structural entity status: the more often an entity provides the anchor point of a theme (a contextual link), the more central it is for linking the discourse together, the more pivotal to discourse structure, and the more likely to be part of the gist of the discourse.

3.4.2 Discourse Topic

In Section 3.4.1, I have argued that on the sentence level, thematic constituents link sentences to the preceding co-text. But as we have seen in Section 3.3, there is more to discourse than just cohesive, grammatical sequences of words. The main criterion for calling an utterance or a sequence of utterances a discourse is that they serve a common discourse purpose. How does this discourse purpose relate to sentence themes? And what about discourse-level topics? These are the two questions we will discuss in the following pages.

Discourse topics are what a piece of discourse is about. This intuition has been explored in several different ways. Scholars who want to link sentence- and discourse-level topics tightly would probably want to express the topic of a discourse by a noun phrase. The discourse topic (or discourse segment topic) would be a special discourse entity among others, maybe with a flag set in the set of structural status variables. A good example for such a strategy is (Dik 1989, Section 13.3).[8] Discourse topics are "those entities about which a certain discourse imparts information" (Dik 1989, page 267). Discourse topics may be more or less central to the discourse. They are hierarchically organised. For example, the topic of the present section, "discourse topic", is but a subtopic of the topic of Section 3.4, "theme and topic". Once a topic has been introduced, it is given. Levy (1982) proposes to measure whether a discourse entity is in fact discourse topic on the basis of its co-specification sequence. Her main results are reproduced in Table 3.1. This proposal comes very close in spirit to what I will argue for in Chapter 5.5, but I would not go as far as to claim that what I am measuring is topicality. She found that topical discourse entities tend to be introduced in non-subject position by a long, very explicit NP, while subsequent mentions of that entity show exactly the reverse pattern: attenuated forms, such as pronouns, and a preference for the subject position.

A notion of discourse topic restricted to NPs would of course fit in perfectly well with our structural entity status. But there is more to the intuition that discourses are about something than special discourse entities. What the addressee perceives to be the main content, the main intention of a discourse will influence how he processes it. The topics of many texts are never verbalised in the text itself, much less in NP form. Of course, in English and German, every complex proposition can be transformed into an even more complex noun phrase—although criticsm of the oft-used and oft-misused (mostly by politicians) possibility of the transformation of preposterous propositional complexes into nonsensical noun phrase monsters parsable

[8] Another excellent example, (Givón 1992), will be discussed later, in Section 4.3.5.

criterion	less	more
	likely to be discourse topic if	
length of co-specification sequence	shorter	longer
mean distance	greater	smaller
first mention		
position	subject	predicate
form	underdetermined	overdetermined
subsequent mentions		
position	predicate	subject
form	overdetermined	underdetermined

Table 3.1. Criteria for Determining the Topicality of a Discourse Entity. Adapted from (Levy 1982, Table 2, page 301)

only by trained linguists with considerable experience in the construction of complex unstarred material for research purposes by elderly guardians of the language is frequent. So, if we still want to introduce a discourse entity for the discourse topic, we cannot presume that this entity will be mentioned in the discourse itself, but that it has to be reconstructed from the discourse by rules. This is the reason why topicality cannot be established quantitatively. Discourse topics can only be establishing by *interpreting* the text at hand.

But this leads us to another problem: each text has several possible topics, and several possible summaries. The problem persists when we switch to an intentional theory of discourse structure. In order to determine how a speaker chose to structure a discourse in order to reach the goal of his plans of speaking, we need to interpret that discourse, sometimes in considerable detail, and in the end, everything will be merely guesswork. van Dijk (1985a, page 77) summarises the problem as follows: "Thus, we may need complex social and political knowledge schemata, or *scripts* to understand what this text is about (Schank 1977, Schank 1982)." I will discuss three proposed solutions here, one which was largely developed on written language (van Dijk), one which relies on general pragmatic principles (Wilson) and one which was developed using spoken language (Chafe).

Deriving Topics Structurally: Van Dijk defines discourse topic on the level of semantic macrostructures, in which the propositions that code the content of a text are organised (c.f. page 28f.) Such macrostructures are motivated by the observation that people can usually tell what a discourse or a discourse segment is about, and that they can give a summary of the main points. van Dijk (1980) wants to model this ability by a set of *macro rules*. Ideally, analysts should be able to derive the topic of a text directly from its propositional coding. The tools for this derivation are a set of four macro rules, which are summarised in Table 3.2.

It follows from van Dijk's procedure that the topic of a discourse does not need to be mentioned explicitly in the text, especially if it is derived by generalisation and integration. He makes this point very explicit in (van Dijk 1981), where he discusses sentence and discourse topics and comes to conclusions which are very close to those I argue for in this section.

Although van Dijk's approach can be applied to text analysis, Gülich and Raible (1977, page 272 ff.) observe several problems with it. Firstly, it is difficult to derive propositions from texts, and van Dijk is not very explicit about how this should be done. Despite these

OMIT:	omit all unnecessary information, all information that will not be relevant later in the text
SELECT:	omit all propositions that are included in, presupposed by, implied by, or conditions for, other propositions
GENERALISE:	abstract away from certain properties of referents
INTEGRATE:	summarise a sequence of propositions

Table 3.2. Van Dijk's macrorules for topic extraction (van Dijk 1980, van Dijk 1985a)

problems, Früh (1992a) developed a content analysis method called "Semantische Struktur- und Inhaltsanalyse" (Semantic Structure and Content Analysis, SSI) which takes a detailed propositional coding of source texts as input. Second, Gülich and Raible point out that it is not at all clear when which macro rules should be used. Van Dijk himself acknowledges that problem. He admits that summaries can be very subjective (van Dijk 1985a). But this does not lead him to reject the idea of macrorules. His solution is different: In his later work, he interprets his macrostructures, superstructures and macrorules not as fixed schemata. Macrostructures are constructed dynamically. Superstructures, which code text-type specific schemata, move towards flexible scripts which help the addressee predict how the discourse will develop. Van Dijk's concept of discourse topic also changes:

> [. . .] themes or topics are *cognitive* units. They represent how the text is understood, what is found important, and how relevancies are stored in memory. [. . .] Finally, the cognitive nature of macro-interpretation also requires a more process-oriented approach to the assignment of topics to a text. Whereas abstract macrorules derive topics from a *given* text, or rather from its underlying sequence of propositions, this is not what a reader actually does. [. . .] readers use expedient *macrostrategies* for the derivation of topics from a text.
>
> (van Dijk 1985a, page 76 f., italics in the original)

It is clear that procedures such as those of van Dijk, which operate only on propositions, can fail to capture what a text, especially a literary text, is really about (Gülich and Raible 1977). But for most purposes, especially for the analysis of primary content as defined by Ungeheuer (1967/1972a), his approach is certainly sufficient. Such a semantic analysis also satisfies the intuition that when we ask the question "what was this text about", we do not want to hear speculations about the writer's intentions. The Gemayel text, reproduced in Appendix 6.4.3, is about the reactions to the assassination of Lebanese president-elect Bashir Gemayel in 1982. An intentional analysis would cut deeper, to the levels of Ungeheuer's secondary and tertiary content. What is the intention of the paragraphs? What other intentions (or discourse segment purposes) apart from informing about facts are there?

Once we have accepted the limits of a structural approach such as van Dijk's, once we do not expect the macro rules to operate automatically anymore, but can use them to describe how we would summarise the gist of a text, then Structure Theory is still a valuable research tool, especially since it offers a nice vocabulary for formal descriptions.

Topics and Activation: Because van Dijk focused on the analysis of written discourse, it was relatively easy to posit something like structural rules for deriving discourse topics. If his

primary data had been spoken language, and in particular conversation, his task would have been far more difficult. Researchers in the field of conversation analysis often do not care about formal definitions of discourse topic. For determining what a discourse is about, they rely on their intuitions. Instead, conversation analysts concentrate on ways in which topics are negotiated, maintained, and shifted (Bergmann 2000). Prime examples for this can be found e.g. in (Sacks 1995). To give a complete overview of the notion of "topic" in the analysis of spoken language would lead too far afield, especially since spoken language data will not be discussed further in this thesis. In order to illustrate the kind of approach that appears to be appropriate for speech, I will discuss two activation-based approaches to topic: the topic framework of Brown and Yule (1983) and the aggregate topics of Chafe (1994).

For Brown and Yule, the topic, that which a discourse is about, is an important category of analysis. If we can identify the topic of a discourse, it means that this discourse coheres enough to be more than just an arbitrary collection of utterances. Since there are many possible paraphrases of a discourse, the analyst should not single out one of them as "the" topic. Instead, she should provide enough information for deriving all of the relevant paraphrases. This leads Brown and Yule to replace the notion of topic by that of *topic framework*.

> Those aspects of the context which are directly reflected in the text, and which need to be called upon to interpret the text, we shall refer to as *activated features of context* and suggest that they constitute the contextual framework within which the topic is constituted, that is, *the topic framework*. (Brown and Yule 1983, page 75)

This definition is much closer in spirit to the psychological subject of Wegener (1885) than to van Dijk. Basically, what Brown and Yule propose to do here is to describe the exposition of a complete discourse fragment, that which needs to be activated in the hearer's mind so that he can interpret the discourse.

Chafe (1994) also describes topics in terms of activation, but he takes a rather different approach. His perspective is not that of the hearer or analyst, but that of the speaker. For him, topics are aggregates of

> . . . coherently related events, states, and referents that are held together in some form in the speaker's semiactive consciousness. A topic is available for scanning by the focus of consciousness, which can play across the semiactive material, activation first one part and then another until the speaker decides that the topic has been adequately covered for whatever purpose the speaker has in mind. (Chafe 1994, page 121)

Like van Dijk, Chafe defines discourse topics on the level of content, not on the level of intentions. Chafe's discourse topics are complex structures. Notably, the speaker decides which events, states, and referents she needs to aggretate. Discourse topics are in turn embedded in more complex hierarchical structures. There can be supertopics and subtopics.

3.4.3 Summary

In this section, I have drawn a sharp distinction between sentence themes and discourse topics. Many functions have been proposed for sentence themes; the most fruitful one for our purposes

is that of *contextual link*: if a sentence has a (marked or unmarked) theme, the thematic constituent shows the addressee how to link that sentence to the preceding discourse. This is easy if the theme is a discourse entity which has already been established, and even easier if that entity is salient.

No wonder that givenness plays such a large role in determining possible themes. But these thematic discourse entities need not be established explicitly by referring expressions. If they stand for propositions, it may be sufficient to just mention that proposition once.

In the following example, the theme is an event, the explosion of the laboratory of Unseen University.

(3.7) Suddenly the laboratory burst into flames. The windows splintered; everybody in the courtyard ran to seek shelter. The colours of fire painted hideous pictures of demons against the pitchblack sky.

If communicator and addressee know each other very well, the discourse entity need not even be explicitly mentioned. In the following example, two very nice and discreet colleagues, A and B, are discussing the behaviour of another, completely disgusting colleague C, in a meeting. The first exchange is about C's habit of picking her teeth, while the second adjacency pair is about the proposals she made during a meeting, and the way she delivered one of them.

(3.8)
A: Do you agree, my dear B, that people who pick their teeth in meetings . . .
B: . . . are not very well educated? Yes, I do.
I mean, some of the proposals were quite startling
A: Any proposal that is made through clenched teeth should be rejected outright.

Discourse topic, on the other hand, cannot be reduced to single discourse entities as easily as sentence themes. Following Chafe and Brown/Yule, the topic of a discourse segment is not only what that segment is intended to be about—it also consists of closely related concepts and contextual features that are easily accessible because of their close connection with the topic. These discourse-level topics influence how a discourse is organised, how discourse models are constructed, they activate expectations. They evoke semantic scripts (what will be talked about in the context of this topic?) and pragmatic scripts (how should we talk about it?). The topic of a discourse can always be identified on a metalinguistic level, and spontaneous dialogues are full of instances where topics are negotiated; sometimes, it is in these negotiation subdialogues that topics are made explicit for the first time.

To sum up, while sentence themes should be discourse entities, discourse topics can become discourse entities. Hence, in structural entity status, we should protocol how often and when a discourse entity has served as theme, and describe the connection between a discourse entity and the discourse topic, if it can be established.

3.5 Summary

What *is* structural entity status? How can we describe the role that a discourse entity plays in a given discourse? This has been the guiding question of the extensive discussions presented in this chapter. The short answer is: It depends. The long answer is: Structural entity status is explicated by the answers to three questions:

1. *In which discourse segments does the entity occur?*
   Different theories cover different aspects of the answer. For example, RST provides a very detailed description of the structural relations between discourse segment, as does the theory of van Dijk. The theory of Grosz and Sidner (1986) is more specific about how the mentions of discourse entities in a segment are protocolled. Recently, Cristea et al. (1998) have also extended RST in this direction with their Veins Theory.

2. *How is the entity linked to other entities in the discourse?*
   Most theories are silent about this aspect. The only good way of dealing with that question are the topic framework in the style of Brown and Yule (1983) or Chafe's (1994) discourse topics, which show where potential bridging links to other entities can stem from.

3. *How closely is the entity connected to the speaker's communicative intentions?*
   This question can only be answered by a theory of intentional structure, such as that of Grosz and Sidner (1986).

The discussion in Section 3.4, has shown that we cannot expect that the topic of a discourse will be verbalised explicitly by a referring expression some time during speaking—especially not when we adopt the definitions of Chafe (1994) or Brown and Yule (1983). Although the topic of a discourse may be available for the occasional discourse deictic reference, as in the classic flash enlightenment after brooding long hours over a difficult paper "So *that* was what it was all about!", we cannot take that for granted. What we *can* determine is whether the discourse entity is a central referent in one of the discourse segments, maybe calculated along the lines of Levy (1982), and when it has acted as *contextual link* between two sentences, in other words, when it has been the *theme* of a sentence. The more often a discourse entity occurred in a stretch of discourse, and the more often it served as a contextual link, the more fundamental it will be to the discourse model of that passage.

But thematicity is not the proper level on which structural entity status should be defined. Although theme is useful in describing the role an entity plays in discourse, it is not quite the fundament we want. Rather, the concept needs to be explicated by a theory of discourse structure. In our review of several theories of discourse structure in Section 3.3, we found that although structural entity status can be explicated in all of them, neither is comprehensive enough to provide a proper foundation. The theory of Grosz and Sidner (1986) involves a dedicated level of intentional structure, but it does not provide descriptions for the semantic macrostructure of a text. The theory of van Dijk, on the other hand, bridges psycholinguistic theories of discourse comprehension and linguistic approaches to discourse. With his superstructures, van Dijk provides a formalism for describing genre-specific aspects of discourse structure. Which theory you choose ultimately depends on the research you are aiming for. Because it requires a very fine-grained propositional input, van Dijk's theory is completely unsuitable for all computational linguistics applications which strive to cover large amounts of data. But together with Kintsch's Construction/Integration model, it can be used to specify the input texts to psycholinguistic experiments or running small-scale simulations on toy discourses.

No matter how we explicate structural entity status, we should not make it into a pillar of textual coherence. Modern text linguistics (Nussbaumer 1991) and modern psycholinguistics (Sanford and Garrod 1994) agree that coherence is constructed in the mind of the addressee.

Referential continuity surely helps in constructing a coherent representation of the discourse—after all, discourse models are supposed to be organised around discourse entities—but, as e.g. Givón (1995a) notes, there are many other levels on which coherence can be established, not least spatially and temporally.

4 The Management Dimension

The status of a discourse entity not only reflects the role that it plays in a discourse. Entity status also provides information that is necessary to manage a discourse entity. These terse statements suscite a number of questions which I will try to address in this chapter:

- What sort of management information should entity status provide? (Section 4.1)
- How does that information relate to the procedures people use to construct, access, and update discourse entities? (Section 4.2)
- How have previous researchers modelled the management of discourse entities? (Section 4.3)

In contrast to Chapter 3, where I strived to be as comprehensive as possible, this chapter focuses on one particular perspective of looking at discourse processing, the cognitive perspective, which is invoked over and over again in the literature.

4.1 What is the Management Dimension?

In this section, I survey the main issues involved in managing discourse entities. We can talk about the management of discourse entities in terms of algorithms and data structures, or in terms of cognitive models. I have chosen the more neutral language of algorithms here. The section is divided into two parts. We begin with a brief overview of the empirical phenomena that the management of discourse entities has to deal with (Section 4.1.1). On this basis, I then define in Section 4.1.2 the management operations that entity status needs to support.

Note on Terminology: The terminology in the field of research that we are about to enter is as muddled as in all fields with a long enough research history. Unsurprisingly, the term "anaphor" has three very distinct meanings (Hoffmann 2000). In rhetorics, it is a figure of repetition, in Government and Binding theory, it designates certain types of pronouns, such as reflexives and reciprocals, and finally, in discourse studies, it means a device for pointing back in a text; the favourite such device studied in the field is the common pronoun.

Here, we focus on the third meaning: anaphoric expressions are expressions that point back to something that has been mentioned already in the preceding co-text. Some researchers distinguish between anaphoric and deictic expressions: anaphoric expressions maintain the activation of a discourse entity, while deictic expressions refocus the attention of an addressee on an entity

(Bosch 1983, Ehlich 1982). I will use the term "anaphors" for both types of expressions. More precisely, *anaphors* (singular: anaphor) are all expressions that[1]

1. point to a stretch of discourse in the preceding co-text. The stretch of discourse that an anaphor points back to is called its *sponsor*.

2. that stand in some sort of relation to its sponsor. If the sponsor is an NP and both the anaphor and the sponsor co-specify the same discourse entity, then the sponsor is called the *antecedent* of that anaphor.

Sequences of co-specifying expressions form *co-specification sequences*. [2] In order to distinguish between discourse entities that are only mentioned once and entities that are part of a proper co-specification sequence, I will call the former *deadend* and the latter *tracking*, using the picturesque terms of Biber (1992).

4.1.1 The Linguistic Domain

From a linguistic point of view, the referring expressions that a communicator uses should help the addressee

- establish the correct referents and assign the sentences the correct truth-conditional interpretation (to the extent that it is relevant to successful communication).

  This aspect has been investigated in great detail by semanticists, in particular those who work in the framework of dynamic semantics (Kamp and Reyle 1993, Heim 1983). For a recent introduction see (Heim and Kratzer 1998, esp. Chapter 9). Bosch (1983) surveys some relevant data.

- construct the text as a coherent whole and identify all relevant communicative acts.

  On the neo-Gricean side, Levinson (1987, 1991) and Huang (1993) have probed whether syntactic Binding Theory can at least partly be replaced by conversational implicatures. Dale and Reiter (1995) base their algorithm for generating referring expressions on the Gricean maxims. In the framework of Relevance Theory (Sperber and Wilson 1995), an early paper is (Kempson 1988); more recent work includes (Breheny 1997, Figueras Solanilla 1998).

To dwell on these two tasks further would lead us too far afield here. Instead, let us discuss some of the problems that can occur when people need to perform them on the basis of linguistic data generated by others. I do not aim to survey the solutions that have been proposed, as well—this would lead us too far afield here. I focus on two aspects: constraints on the form of referring expressions other than entity status, and the relation between the anaphor and the co-text that sponsors it.

[1] This definition owes a lot to discussions with Donna Byron.

[2] Since expressions that specify discourse entities do not necessarily refer (c.f. Chapter 2), I could have renamed referring expressions "specifying expressions". I refrained from that step because I regard it as unnecessary terminology overload.

Constraints other than Entity Status

The most basic constraint is exerted by the linguistic options of a given language: whether it distinguishes between stressed and unstressed pronouns, whether it has articles, whether it allows to drop pronouns, whether it distinguishes grammatical gender, whether it indicates switch reference, and so forth. For a typological overview of pronominal systems, see Wiesemann (1986). This means that algorithms which map linguistic forms onto instructions for managing discourse entities will differ depending on the language they have been developed for. How large that difference will have to be, if it amounts to a mere flick of a parameter setting, as generativists would have it, or whether it requires deep-seated changes, is an open question; it can only be answered by dedicated functional comparative research.[3] To make matters even more complicated, the linguistic means for referring back to something are not limited to pronouns, noun phrases, and nouns. Such expressions can also be adverbs or verb phrases, as Braunmüller's (1977, Sections 1.1.1–1.1.6) taxonomy of pro-forms shows.

In order to determine which principles underlie the bewildering variety of forms and systems, we would need to delve deep into typological studies, which show how pronominal systems can be organised, and into diachronic studies, which show how they evolve over time. From a functionalist point of view, it is tempting to search for these principles in the mechanisms of referring. This is what Ariel (1988) has done. She argues that systems of referring expressions are organised according to a very simple principle: the less accessible the discourse entity, the more phonological material the form of the referring expression needs to contain. Pronouns, which contain very little semantic material (often just gender and number), tend to be very short, while nouns and noun phrases are longer. The longer a NP is and the more modifiers it contains, the more information it can potentially convey. Ariel counts stress as additional phonological material. By this move, she can account for the finding that the antecedents of stressed pronouns are usually farther back in the discourse than those of unstressed ones, which is corroborated by (Givón 1983b). We will come back to Ariel's proposals later on, when we discuss her accessibility theory on pages 80f.. To evaluate her claims about the architecture of pronominal systems properly would lead us too far afield here. Although the amount of linguistic material which needs to be presented to the addressee certainly determines which referring expression can be used when, accessibility alone cannot account for the variation that we see.

Syntax: Avid functionalists vividly deny that some constraints on the form of anaphors can only be expressed via syntax. But exactly that is the central tenet of the Binding Theory of generative grammar (Chomsky 1981, Fanselow and Felix 1987, Sternefeld 1993). The central

[3] In this respect, much research, in particular in formal semantics, is extremely limited, because it adheres to the time-honoured principle to take any language, say, English (or Dutch or German, to be fair, but that does not make too much of a difference). In principle, the resulting formalisms should not be affected too much by the choice of language, but it is tempting to justify certain constructs by claiming that they explain the use of the bare NP or the definite determiner in English, arguably one of the most widely spoken languages of the world, but still only one among thousands. Of course, the very same deplorable limitations apply to the studies that I will present in Chapters 6–7; the data represent educated American English and Standard High German, hardly languages that suffer from lack of attention by researchers. To conduct a typologically more satisfying study was beyond the scope of this thesis, not least because any comparative functional research faces many methodological problems (Chesterman 1998).

observation is that some types of pronouns, such as reflexives, need to be bound by an antecedent in their immediate syntactic vicinity. To give an idea of the machinery that is needed to formalise this observation, I will present a short summary of the classical Government and Binding (GB) binding theory as described by Haegeman (1994, Chapter 4). First, we need a special structural relation on syntax trees, *c-command*, which is used heavily in GB theory.

Definition 4.1 (c-command) *A node A* **c-commands** *a node B iff*
(i) A is not a daughter of B in the syntactic tree
(ii) B is not a daughter of A
(iii) the first branching node that is a mother of A is also a mother of B

Reflexives and reciprocals are *bound*. In generative theory, only these two types of pronouns are called anaphors. Binding is defined as follows:

Definition 4.2 *Binding A* **binds** *B iff A c-commands B and A and B are co-indexed, that is, share the same referent.*

Now, we can formulate the three principles of Binding Theory:

Principle A (anaphora): An anaphor X must be bound in the smallest domain that contains X, the governor of X and either a subject or an abstract agreement element specified for number and gender that occurs in subject position which is co-indexed with X. This domain is also called the **governing category**.

Principle B (pronouns): A pronoun must be **free** in its governing category. Free means 'not bound'.

Principle C (other referring expressions): All other referring expressions must be free everywhere.

In Example 4.1, the subject "The Lecturer in Recent Runes" is co-indexed with the transitive object of the verb "to shave". It is the subject of the major clause to which the verb is associated. That clause is the domain in which the transitive object must be bound, and the only available subject is the Lecturer. If the object of "to shave" is an anaphor, such as a reflexive, Principle A stipulates that it *must* be co-indexed with our specialist in Recent Runes, because the anaphor must be bound in that clause. If, on the other hand, the object is a pronoun, then that pronoun must be free in that clause, and hence, it may not be co-indexed with the Lecturer.

(4.1) [The Lecturer in Recent Runes]$_i$ shaves [himself]$_i$.

These three principles are of course not the last word on Binding Theory, and the extent to which they apply is still debated. For example, Levinson (1991) argues that although Principle A may be truly syntactic, Principles B and C can be motivated pragmatically using general conversational implicatures (GCI). These GCIs posit strong but defeasible defaults for interpreting referring expressions.

Social Constraints: Finally, let me mention some constraints that are diametrically opposed to the Binding Theory we have just discussed—social conventions. As the following example demonstrates, social and stylistic conventions can be so strong that they prevent the uninitiated from resolving pronouns to the correct discourse entity:

(4.2)

Julius Caesar:	Nachdem Vercingetorix von den Galliern geschlagen war, legte er seine Waffen dem ruhmreichen Führer zu Füßen [. . .]
Roman 1 (to Roman 2):	Von wem redet er?
Roman 2:	Von sich. Er spricht von sich häufig in der dritten Person.
Roman 1 (to Caesar):	Er ist großartig!
Caesar:	Wer?
Roman 1:	Na Ihr!
Caesar:	Ach, Er!

from (Goscinny and Uderzo 1971/1974)

Pronouns are used to draw lines between "us" and "them" (Brown and Gilman 1960, Mühlhäusler and Harré 1990). Norbert Elias puts it this way:

> Der Satz der persönlichen Fürwörter repräsentiert den elementarsten Koordinatensatz, den man an alle menschlichen Gruppierungen, an alle Gesellschaften anlegen kann. Alle Menschen gruppieren sich in ihren direkten und indirekten Kommunikationen miteinander als Menschen, die in bezug auf sich selbst "Ich" oder "Wir" sagen, die "Du", "Sie" oder "Ihr" in bezug auf diejenigen sagen, mit denen sie hier und jetzt kommunizieren und "Er", " Sie", "Es" oder, im Plural, "Sie", in bezug auf Dritte, die momentan oder dauernd außerhalb der hier und jetzt miteinander kommunizierenden Personen stehen. (Elias 1970, page 133)

Although the first- and second-person pronouns that Elias mentions are not regarded as anaphoric in most of the literature, they nevertheless specify discourse entities, and their use has social impact. Not only do social structures influence who gets referred to by which referring expressions, there are also class-specific strategies for structuring co-specification sequences. For example, Hemphill (1989) found that working class girls tended to maintain the discourse topic by pronominal anaphora and ellipses across speaker turns. Middle-class girls, on the other hand, tend to restate the topic with a full NP when they mention it for the first time in a turn.

Relations between Referring Expression and Co-Text

There are three basic types of relations between a referring expression and its co-text:

1. it accesses a discourse entity that has already been established in the discourse model,
2. it is linked somehow to other discourse entities in the co-text, such as part-whole relations,
3. it mentions a completely new discourse entity.

On the following pages, we will discuss each of them in more detail.

The Base Case: Identity. The relation between anaphor and antecedent or sponsor is not always straightforward. Even if both access the same discourse entity, their linguistic forms can be related in three fundamentally different ways: morpho-syntactically, semantically, and pragmatically (Braunmüller 1977, Cornish 1986). We have syntactic relations when the semantics of the anaphoric expression is largely determined by the semantics of the antecedent. In Example 4.3, expression 3 stands in such a syntactic relation to expression 1. Cornish explicitly restricts this relation to grammatical morphemes which function as anaphora, such as personal and possessive pronouns. Semantic relations rely on lexical semantic relations between referring expressions, such as hyponymy, hyperonymy, and synonymy. For example, "book" (referring expression 4) is a hyperonym of "grimoire" (expression 2). Cornish files anaphora that repeat the head noun of the antecedent expression, or full NPs that take up an earlier verb, under the category "other anaphora", but if we allow for a richer set of possible semantic relations, one that also covers repetitions and morphological derivations, we can also classify these cases under the heading "semantic relation".

Finally, there are those relations that can only be established by world knowledge. If you have not read Appendix A.2 already, or if you are not familiar with the works of Terry Pratchett, you should have problems in establishing the link between expression 8 ("the ape") and expression 3, which accesses the entity corresponding to the Librarian of Unseen University.

(4.3) [The Librarian of Unseen University]$_1$ gently dusted [the old grimoire]$_2$. [He]$_3$ put [the book]$_4$ back on [the shelf]$_5$. [The cover]$_6$ had been really dirty, but now [it]$_7$ gleamed beautifully. [The ape]$_8$ was pleased with [[his]$_{10}$ work]$_9$.

Suddenly, [the Librarian]$_{11}$ heard [footsteps]$_{12}$. [A student]$_{13}$ emerged from [the darkness]$_{14}$ to ask [him]$_{15}$ a question. [They]$_{16}$ talked for a while, then [he]$_{17}$ went back to work, polishing and dusting.

In fact, writers often use definites in order to recall pragmatic information about a discourse entity. Definite descriptions can be analysed as a function that is applied to a discourse entity (Löbner 1985, Fraurud 1990); this function can predicate properties of that entity which the addressee did not know yet, but has to infer from conventions. Whether that attempt succeeds depends largely on whether the required pragmatic relations are known to the addressee. For example, if you did not know Angua as a werewolf, only as a member of the Ankh Morpork City Watch, and if I were to say to you: "Oh, I saw our little werewolf last night. It was full moon again.", you would be completely puzzled. The following example is another nice illustration, with a scholarly comment by an (initially) puzzled linguist:

> Example:
> *. . . Mr Hart's campaign in Washington has effectively acknowledged that if the senator is to gain the nomination, it will not be in the primaries and caucuses lying ahead.* (The Guardian, 21.4.84, page 8)
> Comment:
> Indeed, when I originally read the article from which (30)a, for example, is taken, I immediately took *the senator* to refer to some other individual than Mr Hart, but quickly revised this interpretation on finding that there was no other named or inferable individual within

> the discourse context which it might coherently describe. It is important to remember, however, that anaphors often indirectly *inform* the addressee of some property of their referent, of which they may previously have been unaware.
>
> (Cornish 1986, Example 30(a), page 24; Footnote 9, page 33)

In general, the expressions that people use to specify discourse entities depend very much on their current situation model and in particular on the set of beliefs that they hold. These problems have also been discussed under the heading "point of view". When resolving referring expressions in a stretch of discourse, it is important to know the "psychological point of view" of the person from whose perspective the discourse is written; an algorithm for resolving such references can be found in (Wiebe 1994).

Bridging: It becomes even more difficult to establish a connection to the co-text when the anaphor is *associative*, in other words, when the anaphor evokes a new discourse entity which is linked to other discourse entities in the preceding co-text (Clark 1977, Asher and Lascarides 1998, Charolles 1999). For example, the NP "the cover" (expression 6 in example 4.3) stands for "the cover of the book". It can only be resolved by somebody who knows that books tend to have covers. Similarly "the shelf" (expression 5): Books tend to be stored on shelves. Such cases have also been called *bridging* or *textual ellipses* in the literature. The term "textual ellipsis" points out the fact that the communicator has 'elided' a term that explicitly establishes the relation between anaphor and co-text. In the remainder of this thesis, I will use the term *bridging* for the phenomenon itself because it is the term that is most often used in the psycholinguistic literature, and *associative anaphor* for the anaphoric expressions that need to be resolved using bridging inferences.

Many scholars have attempted to classify bridging relations into types according the connection between the anaphor and its sponsor. Examples for some typical bridging relations can be found on page 102f.. Clark (1977) argues that such classification may be pointless because they can never hope to be exhaustive. What is important for resolving an associative anaphor is that a bridge between anaphor and a sponsor in the co-text can be found *at all*, not that the bridge is of a specific type. Some types of bridging relations have received particular attention in the literature, either because they touch on interesting semantic problems, or because they can be modelled using relations such as meronymy or hyponymy that are a staple of lexical semantics.

Plurals: Formal semanticists have long been interested in the resolution of plural pronouns. (Kamp and Reyle 1993, Chapter 5) discuss some of the problems with plurals in the light of Discourse Representation Theory. For example, expression 16 in Example 4.3 refers to both the Librarian and the student. A related question is: Under which conditions can we use a pronoun in order to refer back to one of the constituents of a coordination? The pronoun in expression 17 eventually resolves to the Librarian, but only after the rest of the sentence has disambiguated the reference.

Evolving Entities: So far, we have been discussing how discourse entities are initialised (first mentions) and accessed. However, one crucial operation is still missing: updates. A discourse entity can change during a discourse up to the point that it ceases to exist as a separate entity.

The examples for such *evolving anaphors* typically revolve around two question: How do we refer to transvestites (Example 4.4), and how long is a chicken still a chicken (Example 4.5)? In the first example, we can refer back to Julius Maria N. after the revelation that he is female by two pronouns: male singular and female singular. The male pronoun would point to the initial description, the female pronoun is licensed by the recent discovery of his/her (?) true sex. In the second example, the chicken is still available as a discourse entity, even though it has long since ceased to be a coherent entity in real life. Recent research centres on the degree of ontological change (or, in the case of the unfortunate chicken, decomposition) that a discourse entity must undergo before it cannot be referred to by a pronoun anymore (Kleiber 1997).

(4.4) Durch die Belästigung unseres Chemieassistenten Julius Maria N. wurde von unserer Verwaltung festgestellt, daß es sich bei ihm um eine Assistentin handelt. (Wittich 1976/1986, page 52)[4]

(4.5) Kill an active, plump chicken. Prepare it for the oven, cut it into four pieces and roast it with thyme for 1 hour. (Brown and Yule 1983, Example 16, page 202)

First Mentions: At first sight, it appears intuitively clear which referring expressions count as first mentions, and which do not. But when we set out to determine first mentions in data, the picture suddenly becomes blurred. What about coordinations where all of the coordinated NPs have never been mentioned in the discourse before? Do they establish new discourse entities? What about NP modifiers, such as "Firth" in "the *Firth* School"? What about NPs that occur in idioms, such as "a stroll" in "to take a stroll", or "a few more pages" in "to write a few more pages"? These open questions led Behrens and Sasse (1999) to concentrate only on those referring expressions that refer back to an already established discourse entity, discarding first mentions almost entirely. I call this problem the *initialisation* problem: When does a discourse entity become available for anaphoric reference? In dialogue research, this problem has also been analysed under the label of "grounding" (c.f. e.g. Traum 1994, Poesio and Traum 1997).

A particularly interesting instance of the initialisation problem are *anaphoric islands* (Postal 1969, Ward, Sproat and McKoon 1991). Originally, Postal (1969) claimed that anaphors cannot have antecedents that are somehow part of lexical items. He called these items anaphoric islands. In particular, this would imply that it is not possible (or at least extremely difficult) to refer back to a part of a compound or the source of a derived word. But there are plenty of exceptions to this rule. In the following example, the first example is clearly out, while the others have all been claimed to be acceptable, at least idiolectally (Douloureux 1971, Corum 1973):

(4.6) * The blonde got it caught in the fan. vs.
The girl with the blonde hair got it caught in the fan. (Postal 1969, Example 8)
(German equivalent: Der Blondine hat es sich im Fön verhakt.
Dem Mädchen mit dem blonden Haar hat es sich im Fön verhakt.)

(4.7) McCarthyites are now puzzled by his intentions. (Postal 1969, Example 42)
(German equivalent: Kantianer sind mittlerweile von seinen Schriften verwirrt.)

[4]Because of the harassment of our chemistry assistant Julius Maria N., it was discovered by our administration that he is in fact a female assistant.

(4.8) When little Johnny threw up, was there any pencil-eraser in it? (Douloureux 1971, Example 7d)
(In German equivalent, adverb instead of pronoun: Als Hänschen Klein kotzte, war da Radiergummi drin?)

Oakhill and Garnham (1992) had subjects judge the acceptability of some doubtful cases of anaphoric reference, which included anaphoric islands. They report that subjects tended to reject them as poor style. This result partially validates Postal's (1969) initial intuition.

After this excursion onto remote anaphoric islands, let us now return to normalcy—or what researchers have argued normalcy to be. The prototypical first mention comes with an indefinite article (Heim 1983, Kamp 1981). Weinrich (1976, page 168f.) postulates that the indefinite article directs the attention of the addressee to what follows, while the definite article directs his attention to what has come before. This received wisdom has been shattered repeatedly by corpus studies; one of the best-known of these is probably (Fraurud 1990). She found that most of the definite descriptions in her corpus of newspaper articles were first mentions. Furthermore, definites tend to be preferred as first mentions if the discourse entity is the beginning of a co-specification sequence. These findings are corroborated by the corpus studies in Chapter 6 and Appendix C. What could communicators intend with this pattern?

To get an idea of the possible answer, we need to go back to semantic research into the meaning of definites. There is a long-standing controversy between semanticists about whether a definite description can only be used felicitously when its referent is unique, or whether that referent needs to be identifiable (Lyons 1999). The uniqueness hypothesis, which can be traced back to (Russell 1919/1993), states that there is one and only one referent to which the definite description refers. The familiarity (or identifiability) tradition, on the other hand, assumes that the definite description contains enough information for the addressee to identify the referent. I will not attempt to argue for one of the positions here. Let me just spell out what they mean for first-mention definites. On the identifiability hypothesis, these definites provide the addressee with enough information to immediately identify the referent of the expression, to construct a new discourse entity with strong links to the preceding co-text or the larger situation (c.f. also the typology of definite referring expressions discussed in Hawkins 1978). On the uniqueness hypothesis, the referring expression provides sufficient information so that the addressee cannot confuse the discourse entity it evokes with other, already existing ones. Thus, no matter what hypothesis you subscribe to, it intuitively makes sense that many co-specification sequences start with a definite NP—whenever that is possible, of course.

It is common for definite descriptions to occur as first mentions of a discourse entity (Fraurud 1990), but not for pronouns. Such first-mention pronouns have received much attention from psycholinguists (Oakhill, Garnham, Gernsbacher and Cain 1992, Gernsbacher 1991, Carreiras and Gernsbacher 1992), who dubbed them *conceptual anaphora*. Table 4.1 summarises the three types identified by Gernsbacher (1991). These first-mention pronouns are so common that native speakers hardly judge them ungrammatical anymore, although processing is more difficult for those conceptual anaphors that do not refer to collective sets.

Type	Example
multiply occurring item or	I need a plate.
frequently occurring event	Where do you keep them?
generic type	I was really frightened by a Dobermann.
	They are dangerous beasts.
collective sets	Last night we went to hear a new jazz band.
	They played for nearly five hours.

Table 4.1. Types of conceptual anaphora after (Gernsbacher 1991); examples from (Oakhill et al. 1992, Table 1, page 261)

4.1.2 The Processing Domain

In the preceding section, we discovered a veritable bestiarium of referring expressions and their uses. We saw that there are many other constraints on the form of referring expressions than entity status, and we found a bewildering number of relations between anaphora and their sponsors and antecedents. From the overview, three main operations on discourse entities emerge: initialise, access, and update. Each participant in the discourse has her own discourse model based on her Personal Experience Theory; this is the source of the point-of-view problems discussed earlier.

When new discourse entities are *initialised*, the entity is created, and an initial description is constructed on the basis of the referring expression and additional knowledge from memory. Webber (1981) emphasises that it is important to construct cogent initial descriptions. The reason for this is simple: the more an addressee knows about a discourse entity, the better he can identify it when it returns in the discourse. What kind of information do we need for this initialisation procedure?

1. We need to know how much information there is in the referring expression itself and whether we need to retrieve additional information and where we might look for it. Pronouns very likely require special strategies, and definite descriptions should send the algorithm scurrying for a fitting place in the current mental model (Johnson-Laird 1983), especially if they are short. In Accessibility Theory, the form of the referring expression would determine the radius in which potential antecedents are searched.

2. We need to know where and how the new entity is supposed to fit into the discourse model. Does it fill a slot in an activated schema? Does it trigger inferences to establish coherence? How is it linked to other, existing, entities?

The necessary information for the second part, the integration into the discourse model, should come first and foremost from the discourse model itself; the structural entity status of already established entities is a secondary source of relevant data. The first part, determining how much information the referring expression yields and deciding on the processing strategies to try—that is pure management.

When a discourse entity is *accessed*, we need the following information:

- *knowledge about the entity.* Under that heading, I subsume all that we know about the discourse entity, in particular, all that has been predicated about it in the discourse.

- *co-text: related discourse entities.* Since discourse entities are linked via co-occurrence relations and background knowledge, one entity can serve as the access route to another. Such connections are necessary whenever anaphoric and antecedent expression are related neither syntactically nor semantically, as in the following example.

  (4.9) Gerd and Maria, two researchers at the same institute, go for a swim together every now and then. [Gerd]$_G$ swims slowly, but steadily. That's why [Maria's colleague]$_G$ prefers to go swimming when the pool is less full.

  The second expression that refers to Gerd is not a pronoun, although that would be allowed here; instead, the reader has to activate the appropriate connection via Maria.

- *context: related items or knowledge schemata in a long-term store.* For other pragmatic and all semantic relations, such as hyponymy or meronymy, we need to access knowledge that is kept in a long-term store. In Example 4.9, the reader could also have identified the reference to Gerd correctly if he had assumed that slow swimmers don't like full pools.

- *salience.* This variable is computed by many resolution algorithms (for a classic example, see Lappin and Leass 1994). It forms an important part of almost all theories with a cognitive basis. Since there are potentially infinitely many links to co- and context, and since much information can be stored about a discourse entity in the course of a long text or conversation, we should have some kind of salience ordering on those three kinds of information as well.

Finally, we need to *update* our discourse entities, adding new information, new connections to other entities, and changing the activation of both the entity itself and the three kinds of access information that we have identified. The case of evolving anaphors has shown that new, contradictory information about an entity should not necessarily override old information. In other words, we need to trace how and when the entity changes during the discourse.

What I have presented on the preceding pages is not a theory of how discourse entities are managed. Rather, it describes what such a theory should be able to explain—at the very least.

4.2 How do People Process Texts?

Just as I have evaluated several theories of discourse structure with respect to what they say about structural entity status in Section 3.3, I will now review a few psycholinguistic theories of discourse processing to see what they can tell us about how discourse entities are managed. Why this focus on cognitive theories instead of on semantics or pragmatics? First, as I have already mentioned, many theories of entity status (or "givenness") use terms from cognitive psychology, such as long- or short-term memory. Such theories can only profit (and some have already profited) from looking behind the folk-psychological interpretation of these terms. Second, psycholinguistics is an experimental science. Once a theory of entity status has been couched in terms of a psycholinguistic theory, a whole new array of methods becomes available

to test it. Finally, in the part of the computational linguistics literature that speaks of discourse entities, researchers tend to assume that something like discourse entities are not only needed for the computational representation of a discourse. They are also needed for modelling the *mental* representation of a discourse.

The remainder of this section is structured into three parts. First, in Section 4.2.1, I very briefly review the cognitive foundations of discourse comprehension: the structure of memory and the inferences people draw in processing discourse. I then review two approaches to the processes of referring expressions (Section 4.2.2): the Mental Models perspective, and the Scenario Mapping and Focus (SMF) perspective. In Section 4.2.3, I compare what the two perspectives can tell us about the management of discourse entities.

4.2.1 Cognitive Foundations: Memory and Inferencing

Let us start with the most basic question: What is it that we are supposed to retrieve knowledge from and store knowledge in? What is memory? In fact, nobody knows for sure (Glenberg 1997), but there are some fairly standard working hypotheses which are surveyed in great detail by Baddeley (1998). His work will be my main source for the following paragraphs.

There are at least two components of memory: a short-term storage (STS) and a long-term storage (LTS). The content of the LTS is often referred to collectively as *world knowledge* in discourse comprehension research, and short-term storage is referred to as *short-term memory*, although Baddeley argues that this term should be reserved for a certain set of experimental techniques.

Working Memory: The STS acts as some kind of *working memory*: it stores a very small amount of information that helps perform the cognitive tasks at hand, such as discourse processing. Baddeley (1998) proposes that this working memory is modular. It consists of a central executive, which in turn controls a number of subordinated systems, such as a visuo-imaginary scratchpad or the phonological loop. This multitude of subsystems appears necessary since several tasks can run in parallel in working memory. When tasks are similar enough to interfere, or when there are too many tasks at the same time, performance, in particular reaction times, on these tasks falls. The central executive controls and coordinates these concurrent tasks. Following Norman and Shallice (1986), Baddeley assumes that the available capacity is distributed among these tasks by two mechanisms: an automatic contention scheduler and a supervisory attentional system (SAS) that can modify or interrupt behaviour “at will”. The phonological loop is subdivided into an articulatory control process and a phonological store. The articulatory control does not depend on peripheral muscles; therefore it has been said to control “inner speech”, speech at a stadium where motor commands are still merely planned. It is also active in reading comprehension, feeding read material into the phonological store.

When the working memory of a person is small, it will be difficult for her to process discourse (Daneman and Carpenter 1983, Just and Carpenter 1992). The reason is given in the following citation, which at the same time provides a bird’s eye view of the complexities of mental discourse processing:

> The working memory system is supposed to be implicated in text comprehension because of its capacity for simultaneous storage and processing of information. During text comprehension these simultaneous capabilities will be needed because of the requirement to recognise words, retrieve their meanings, and parse sentences, while simultaneously integrating the current sentence with what has gone before, and deriving an integrated model of the text as a whole. (Oakhill 1996, page 79)

The capacity of working memory differs from person to person (Just and Carpenter 1992), so that any simulation of that part of memory will need to include capacity as a parameter. Examples of such a simulation in the context of computational linguistics can be found in (Walker 1993, Cahn 1998).

From this brief review, we can derive several consequences for the management of discourse entities. Firstly, a terminological quibble—it might be better to stop talking about short-term memory and start talking about working memory, keeping in mind the complex structure just described. Second, it does not make sense to go through a text and state that this or that discourse entity will surely be in working memory. Working memory will be strained by many other processes than resolving anaphoric reference. In order to predict what will be in working memory when, we need to model these comprehension processes in more detail, and we need to state the capacity that our simulated addressee is supposed to have. The Construction-Integration model of Kintsch (1988) makes quite detailed predictions about these aspects, which have also been incorporated into a simulation. The main downside of this model is, however, that coding is very elaborate and time-consuming, as the worked example in (Kintsch 1985) shows.

World Knowledge: The short passage quoted from (Oakhill 1996) highlights two further problems in text comprehension: which inferences need to be made at what point, and how does text comprehension interact with the long-term store (LTS)? We will discuss each of these two issues in turn, beginning with LTS.

Just like working memory, the long-term store is by no means a monolithic unit. There appears to be a semantic memory, where knowledge is stored, episodic memory for events, and procedural memory for sequences of actions and skills. This tripartite division roughly corresponds to that proposed by Tulving (1985). Neuropsychological evidence suggests that there is a separate autobiographic memory for the events of one's own life (De Renzi, Liotti and Nichelli 1987). How knowledge is organised is another contentious area. A concept from Artificial Intelligence research that is still popular with psycholinguists is the *schema* (Schank 1977, Schank 1982, Minsky 1975, Norman and Rumelhart 1978). Schemas can be characterised as follows:

1. they have variables that can be filled depending on the context in which they are applied,
2. they can be embedded into one another,
3. they represent experiences, not rules,
4. they are actively used in processing incoming perceptual data, not only in discourse comprehension, but also in face recognition, etc.

When an addressee knows the relevant schemata for a particular discourse, he can process it faster and recall it better (Bartlett 1932, Bransford, Barclay and Franks 1972). Schank (1982) distinguishes between plans, scenes, memory organisation packets (MOP), and thematic organisation points (TOP). TOPs encode high-level structural analogies between two situations. An example for a TOP is the analogy between writing a thesis and having a child. Both processes can take a long time and be very painful. *Plans* couple sequences of actions to motivations and goals. *Scenes* associate goals and actions with typical settings. Scenes can be hierarchically organised into sub-scenes; they can be very specific or very general. Batches of scenes that keep co-occurring are organised into MOPs, which might again be organised into meta-MOPs. Kellerman, Broetzmann, Lim and Kitao (1989) apply the concept to discourse analysis. Using recorded dialogues, they identify the scenes that contribute to the MOP of "getting acquainted". Some of the scenes they identified were 'greeting', 'health', 'introduction', 'education', or 'hometown'. Whether we should assume that there are fixed MOPs and schemas, or whether these structures should rather be viewed as emergent, is still hotly debated (Baddeley 1998, Eysenck and Keane 1995).

Inferences: How addressees use world knowledge in understanding a discourse is a wide open research question (Singer 1994). Graesser and Kreuz (1993) identify eleven types of inferences, which are reproduced in Table 4.2. Referential inference is the first and most basic inference readers can make. It is right at the top of the table, and yet, as we have seen on page 62, the connections between anaphors and their antecedents are so rich that researchers despair of ever developing an adequate taxonomy of them.

Graesser and Kreuz (1993) subscribe to a constructionist view on discourse comprehension. When addressees process a discourse, they construct a model of it (a situation model in the tradition of (van Dijk and Kintsch 1983, Kintsch 1988), a mental model in the school of (Johnson-Laird 1983)), and in order to construct that model, they need inferences from world knowledge, both on-line, while they read or hear the discourse, and off-line.

But these inferences are not drawn automatically. Addressees come to texts with a specific purpose in mind, and that this purpose determines how they will read the text and which inferences they will choose to draw. What addressees know about genre conventions or specific text types such as narratives or expository text comes into play here, because that determines how they will expect the text to be structured. Work on the comprehension of narratives has shown, for example, that the deeds and misdeeds of primary characters persist longer in memory than those of secondary characters (c.f. the reviews of Sanford and Garrod 1994, van den Broek 1994).

The complex interaction between reader purpose and recall has hardly be studied yet; no wonder, since most experimental setups require subjects to read pointless texts for no particular purpose. Graesser and Kreuz (1993) make precise predictions about which inferences would be made in which experimental setup. For example, a reader who reads a narrative such as a short story would draw inferences of classes 1, 2, 6, 9, 10, and 11 online, and the others offline.

The problem with a constructionist approach such as that followed by Graesser and Kreuz (1993) and indeed many other psycholinguists (van Dijk and Kintsch 1983, Glenberg, Meyer and Lindem 1987, Glenberg, Kurley and Langston 1994, Fletcher 1994) is that they assume the situation model to contain many inferences from the text, where all informations that are not explicitly stated in the text are inferred. But who decides which inferences are made when, and

No.	Type of Inference	Inference is . . .
1	*referential*	co-specification with antecedent or sponsor in the text
2	*causal antecedent*	causal chain between current action, event, or state and co-text
3	*causal consequence*	forecasted causal chain
4	*instrument*	instrument used when agent executes intentional action
5	*instantiation of noun category*	sub-category or exemplar of mentioned noun
6	*superordinate goal*	goal that motivates agent's action
7	*subordinate goal*	goal, plan, or action that specifies how agent's actions are achieved
8	*state*	ongoing state; can include beliefs, knowledge, or personality traits of agents, properties of objects and concepts, and spatial locations
9	*thematic*	main point of the text
10	*emotion of reader*	emotion reader experiences
11	*author's intent or attitude*	author's motive in writing a text and attitude towards the text, its moral, or its content

Table 4.2. Types of Inferences in Discourse Comprehension. After (Graesser and Kreuz 1993, Table 1, pages 148–149)

when inferencing is supposed to stop? McKoon and Ratcliff (1992) propose a radical way out of this quandary: readers only make those inferences on-line, that is while reading a sentence, which require minimal effort. These are all inferences that are necessary to establish local coherence, such as those that are needed to establish co-specification sequences, and inferences that come from general knowledge, such as "a collie is a dog". They do not expect inferences that are required to establish global coherence to occur except when they suit the readers' purpose and are obvious.

This so-called *minimalist hypothesis* has been hotly debated for several years now (Trabasso and Suh 1993, Foertsch and Gernsbacher 1994, Sanford and Garrod 1998). That controversy is extremely important for the management of discourse entities. If we assume that some sort of mental models are constructed, then they will exert powerful constraints on how new referring expressions are to be interpreted and provide standard access routes to discourse entities at particularly prominent positions of the model. A second consequence is that global coherence will become more important, since to build a mental model of a discourse requires that it can be interpreted as a (somehow) coherent whole. If, on the other hand, local coherence, which is the domain of e.g. Centering Theory (Grosz et al. 1995), is as central as McKoon and Ratcliff (1992) claim, linguistic theories of the processing of referring expressions need to rethink which linguistic structures are supposed to play a part in local coherence, and to what extent discourse structure information is really necessary.

The brief survey in this section has demonstrated that if linguists want to couple their theories about the management of discourse entities with theories about how discourse is processed, they cannot just retreat to "a" standard theory, but need to choose between competing theories

that make different predictions about what is important for anaphora resolution.

In the following section, I present two psycholinguistic approaches to anaphora resolution, the Mental Models perspective to discourse comprehension, and the Scenario Mapping and Focus (SMF) theory of Sanford and Garrod (1998). The Mental Models perspective was chosen because it represents a true blue constructionist account, while SMF has adopted elements of both the minimalist and the constructionist perspective. I have intentionally left out the construction-integration model of Kintsch (1988). Because of its tight links with linguistic theory, it is discussed *passim* here. Its foundations were discussed in Section 3.2.2, and a linguistic theory of referring expressions that is compatible with it will be outlined below, when we discuss the work of Talmy Givón in Section 4.3.5.

4.2.2 Theories of Processing Referring Expressions

The Mental Models Perspective: Mental models are structured mental representations that have been used to theorise about a variety of cognitive processing tasks, from reasoning (c.f. e.g. Johnson-Laird 1983, Johnson-Laird, Byrne and Tabossi 1989, Johnson-Laird, Byrne and Schaeken 1992) to discourse comprehension (c.f. e.g. Garnham and Oakhill 1992, Garnham 1996). When addressees interpret a discourse, they construct a complex mental model, and discourse entities are entities within that model. Mental discourse entities are tightly linked to referents in a possible (real or imaginary) world, because mental models are generally seen as *representations* of such a world. New utterances are integrated into the model incrementally. They are first transformed into a propositional representation, and then integrated into the mental model of the current discourse by a system of rules (Johnson-Laird 1983). Mental models are by no means stable or complete; indeed, they change continuously and are often defective.

Mental Models theory predicts that the preferred antecedents of pronouns are strongly influenced by knowledge-based inferences. For example, with verbs such as “to blame”, if X blames Y, then Y will have done something to anger X. Hence, in a sentence like that in Example 4.10, the preferred antecedent of the pronoun is Bill, not John (Garnham, Oakhill and Cruttenden 1992, Garnham, Traxler, Oakhill and Gernsbacher 1996)

(4.10) John blamed Bill because he . . .

The models also direct the kind of inferences from world knowledge that will be made during processing. A text is (globally) coherent if the addressee can construct a mental model for it. Hence, inferences are directed towards building a specific, coherent model, and elaborative inferences that do not contribute immediately to that task should not occur. This implies that when a new discourse entity is integrated into the mental model of a discourse, it is integrated as tightly as possible into the existing representation, and the better and the more detailed the model, the more easily a discourse entity will be accessed. When information about a discourse entity is updated, this update affects all other relevant aspects of the model as well, so that global coherence is not lost.

The more a person knows about a certain subject matter, the more detailed her models are, the more quickly she can make the required inferences to process a discourse (Tardieu, Ehrlich and Gyselinck 1992). This facilitatory effect even occurs when she is presented with input for another modality, pictures that establish a mental model via the visual channel (Glenberg et al. 1987).

minimalism	Text	→	Proposition			
Mental Models	Text	→	Proposition	→	Mental Model	
C/I model	Text	→	Proposition	——————→		Situation Model
		⟶	World Knowledge	⟶	Situation Model	
SMF	Text	→	World Knowledge	[→	Mental Model / Propositions]	

Figure 4.1. Comparison of SMF theory to other models of discourse comprehension. Adapted from (Sanford and Garrod 1998, page 168 ff.)

Stevenson (1996) discusses how pronouns are resolved in a Mental Models framework. Pronouns are in general resolved to the focused slot of the mental model of discourse. This focused slot can be thought of as the backward-looking centre of Centering, which is the current centre of attention and links a sentence to the preceding co-text.[5] The focus is determined by two types of processes: top-down processes such as first mention preference. Here, first mention means that a discourse entity is the first to be mentioned in a sentence. (Gernsbacher, Hargreaves and Beeman 1989) or thematic role preferences (Stevenson, Crawley and Kleinman 1994), and bottom-up processes such as parallelism or connectives. Both strategies interact dynamically. Top-down cues initialise the focus. When bottom-up search cues occur in the text, or when the main verb has been found, which enables the addressee to use further top-down cues such as thematic roles, the focus shifts again, and so on.

The Scenario Mapping and Focus (SMF) Perspective: The Scenario Mapping and Focus (SMF) account of discourse comprehension was developed by Simon Garrod and Anthony Sanford, two researchers who have published extensively on anaphora resolution (Sanford and Garrod 1981, Garrod and Sanford 1989, Sanford and Garrod 1989, Garrod and Sanford 1994, Sanford and Moxey 1995, Sanford and Garrod 1998). This summary is based largely on the recent (Sanford and Garrod 1998).

Contrary to other approaches, which posit that text is translated into an intermediate propositional representation (Kintsch 1988, McKoon and Ratcliff 1992, Garnham and Oakhill 1992), Garrod and Sanford assume that incoming discourse is directly mapped onto world knowledge. Figure 4.1 compares the proposed processing steps for Minimalism, Mental Models, and the Construction/Integration model to the steps proposed in the SMF account.

The linguistic input is processed incrementally. Words and phrases which are (somehow) linguistically salient are processed more deeply than others (Barton and Sanford 1993). The integration process is guided by bottom-up information from linguistic form, such as focusing constructions, and top-down scenarios, which are to a large extent derived from verb semantics. Knowledge-based inferences can only be triggered by the currently relevant scenario(s). This greatly limits the number of possible inferences and makes the model psychologically more plausible. Sanford and Garrod (1998, page 186) assume that skilled communicators provide sufficient cues to the required scenarios early on. Scenarios are retrieved from memory using a fast, passive process. "Passive" means that retrieval proceeds automatically. Goal-directed inferences appear to come in later, after the discourse has been processed in a first pass on the

[5] Centering is discussed in some more detail in Section 4.3.2 below.

	text-based memory	**knowledge-based memory**
dynamic	explicit focus	implicit focus
static (relatively)	memory for discourse	world knowledge and scenarios

Figure 4.2. The Memory Model of SMF Theory. Adapted from (Sanford and Garrod 1998, Figure 1, page 162 and Figure 2, page 163)

basis of automatically activated knowledge. This later stage is where mental models can be built, and where for example alternative interpretations of complex quantified sentences can be explored (Sanford and Garrod 1998, citing unpublished work by Sanford and Moxey).

The SMF model assumes that discourse entities and scenarios relevant to processing the current stretch of discourse are stored in a *dynamic* part of memory, while knowledge about the world and about potential scenarios as well as a representation of the previous discourse are kept in a relatively static memory. Figure 4.2 illustrates this partition. The dynamic part is divided into an *explicit* and an *implicit* focus. The implicit focus contains verb interpretation schemata and frames. It does not have a limited capacity and belongs to the currently activated long-term memory. Explicit focus, on the other hand, is restricted to those discourse entities that have not been integrated into a scenario yet. These entities are kept in working memory; capacity for them is limited. Together, implicit and explicit focus represent the current state of discourse processing. Discourse history is represented as a trace of connections between the sets of implicit and explicit foci that have been used to process the discourse so far.

4.2.3 Comparison and Evaluation

In this section, I have taken a somewhat unusual approach to reporting psycholinguistic results on co-specification. Most authors (a prime example is Ariel 1990, Chapter 0.3) cite an impressive array of results, and integrate these results into the data that their theory needs to account for or be compatible with. There is nothing to be said against this procedure.

Psycholinguists, on the other hand, search the space of linguistic theories for adequate models of language processing. Gordon and Hendrick, for example, recently tried to marry Discourse Representation Theory (Kamp and Reyle 1993) and experimental results on referent processing (1998, 1997). Centering theory (Grosz et al. 1995) is another good example for a cooperation between linguists and psycholinguists. Centering predicts that if its rules are violated, texts will appear less coherent. Psycholinguists have then operationalised this criterion (here: coherence), such as the less coherent a text, the longer it will take people to read it (Gordon, Grosz and Gilliom 1993, Hudson-D'Zmura and Tanenhaus 1998). Centering also suggests that the backward looking centre corresponds to the current focus of attention. When the center of attention is manipulated systematically in a production experiment, the referring expressions people use should conform to that prediction (Brennan 1995, Brennan 1998).

Such an exchange between linguistics and psycholinguistics can also go in the other direction: Given that many linguistic theories of referring expressions and givenness need to fall back on some notion of memory, some restrictions on processing, some model of communicator and addressee, why not take them directly from psychology, instead from commonsense intuition? This is, if I have interpreted his recent publications (Givón 1995b, Givón 1995a, Dickinson and

Givón 1997) correctly, the position of Talmy Givón.

So, what can we learn from Mental Models and SMF about how discourse entities are managed? In both approaches, discourse entities are tied as closely as possible to the mental model that corresponds to the text or, for SMF, to the current scenario. Anaphoric expressions that require semantic and pragmatic links can be resolved by tight links to world knowledge. Links to the co-text are established within the current mental model (Mental Models), or within the web of past explicit and implicit foci that Sanford and Garrod posit as a long-term discourse representation. Sanford and Garrod do not assume that mental models need to be constructed, although they do not exclude that they are built at a later stage of processing, for example when complex quantified sentences need to be represented. Activation is modelled quite elegantly by the SMF theory; their explicit focus, which was inspired by the focus spaces of Grosz, is a set of currently activated discourse entities. The computation of salience in a mental models framework is more complicated; Stevenson models it as an interaction between top-down preferences and bottom-up cues.

As far as I can see, how the representation of discourse entities should be updated when their properties change is still very much an open question. In the Mental Models framework, the complete mental model is updated as necessary (Glenberg et al. 1994). In the SMF framework, Barton and Sanford (1993) have shown that whether a necessary update occurs or not depends on whether the new information is made salient linguistically.

If a new entity is introduced by a definite description, both approaches do not predict any problems, as long as that entity corresponds to a slot in the scenario or mental model or can be connected to it by easy inferences. No separate typology of bridging inferences is necessary, and it is clear why such a typology would be futile, because it would be tantamount to a typology of the roles that discourse entities can play in all scenarios that could ever be conceived. The rich, directed inferences that are available for accommodating associative anaphors are also available to guide access to established entities along the connections already in the model.

Both approaches allow to model that some discourse entities are more salient than others. Sanford and Garrod keep activated entities in explicit focus, while Stevenson shows that it is possible to define a Centering-style focus of attention on mental models once we can define the appropriate procedures for shifting that focus according to top-down and bottom-up influences.[6]

A Mental Models approach has the advantage that it builds on a very general approach to cognitive processing (Johnson-Laird 1983). Input from different modalities is easily integrated in a common model. The SMF theory is attractive in that it combines the parsimony of minimalism with the powerful knowledge-based inferences of mental models. Both approaches also have their weaknesses. Mental Model theory has frequently been criticised as not formal enough, since they are supposed to be *analogical* representations of a real or imaginary world. Other criticisms include that they emphasise top-down information too much. SMF theory also raises a number of questions: Where do the scenarios come from? How fixed are they? Can they be modelled as emergent structures? Linguists who would like to explicate entity status in one of the two frameworks need to consider these criticisms.

[6]Readers who are familiar with the psycholinguistic literature on pronoun resolution will notice that I have given the notion of salience very short shrift here. The reason for this is simple: I am not focusing on the resolution of referring expressions here, but on the role of discourse entities in psycholinguistic theories of discourse comprehension. For a recent survey of relevant literature, see (Arnold 1998, Chapter 1) or (Garrod and Sanford 1994).

4.3 Hierarchies and Taxonomies

In the previous sections, I surveyed the data that theories discourse entity managements need to cover, and highlighted how psycholinguists have approached the problem. Now, I turn to solutions that have been proposed by linguists. These solutions have often been couched in the form of *taxonomies*. These taxonomies are evaluated according to how well they predict which forms of referring expressions should occur in which contexts. As we have seen earlier on page 58 ff., we cannot expect that any single taxonomy will explain all of the variation we find in arbitrary texts. Other influences such as age, social class, and conventions in discourse communities are definitely not negligible. Nevertheless, we can safely predict that taxonomies which relate to the management of discourse entities should explain by far the largest percentage of variation. The reason is clear: If an addressee is to process a discourse quickly and accurately, referring expressions should help him find the entity they specify as quickly as possible.

In the survey presented in this section, I limit myself to taxonomies that are intended to cover all forms of referring expressions. This means excluding for example the taxonomy of definite description uses discussed by Hawkins (1978, 1991). We begin in Section 4.3.1 with a classic linguistically motivated taxonomy, that of Prince (1981), which has been modified further by Lambrecht (1994) on the basis of work by Chafe. Then, I move on to theories that build more or less on the cognitive processing of referring expressions. The first of these is Centering Theory (Section 4.3.2). Next, in Section 4.3.3, I introduce Mira Ariel's Accessibility theory, which models how the accessibility of a discourse entity influences the form of the corresponding referring expression. The implicational Givenness Hierarchy of Gundel, Hedberg and Zacharski (1993), to be discussed in the following Section 4.3.4, is conceived on similar lines. Finally, we move to two scholars who have taken the cognitive foundations of their respective approaches very seriously: Talmy Givón (Section 4.3.5) and Wallace Chafe (Section 4.3.6).

4.3.1 Familiarity

Prince (1981) developed her famous taxonomy of givenness as a reaction to the confusion that surrounds the term "givenness". She categorically restricts her taxonomy to discourse entities. Among the three senses of givenness she surveys, predictability (Halliday 1967), salience (Chafe 1976) and shared knowledge (Kuno 1972, Clark and Haviland 1977), she decides to model shared knowledge in more detail, because it appears more fundamental to her than the others. Her leading question is: how can we assume that knowledge is shared, that something is already familiar to the listener? And, as a corollary, are there gradations of assumed familiarity?

Prince developed a taxonomy of Assumed Familiarity that relies largely on the sources where we can obtain information about the discourse entity that a referring expression specifies. The complete taxonomy is given in Figure 4.3.

If the referent of an expression is co-present, either linguistically in the text or physically in the immediate communication situation, then it is *evoked*. Prince subdivides this category according to the reason for co-presence, textual (Example 4.11.b) or situational (Example 4.11.c). Brown (1983b) suggested introducing a new sub-category of "textually evoked", "displaced". This category covers instances where the last mention of a discourse entity occurred several utterances back, as opposed to in the current or last utterance. The distinction explains to a large extent why some textually evoked entities are specified by a full noun phrase, others by

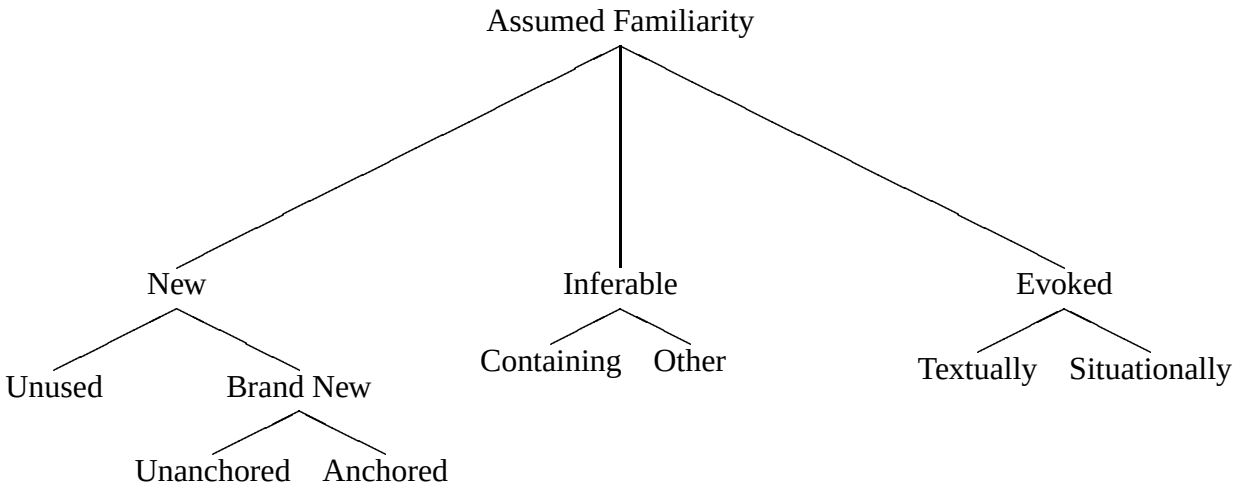


Figure 4.3. Prince's Taxonomy of Assumed Familiarity

a pronoun: Pronouns specialise in short distances to last mention, noun phrases in long ones. In practice, developing consistent schemes for distinguishing between displaced and evoked discourse entities turns out to be difficult.

Inferable (Example 4.11.f) discourse entities can be connected via knowledge-based inferences to already evoked or other inferable discourse entities (hence the name). With her category of *Containing Inferrables* (Example 4.11.e), Prince singles out one of these potential inferences, namely that which connects the member of a set to its superset.

Entities that are completely *new* to the discourse, that have neither been mentioned before, nor can be linked to evoked discourse entities by inference chains, fall into two classes. *Unused* (Example 4.11.g) entities are new to the discourse, but not to the addressee, while *brand-new* (Example 4.11.a) discourse entities are new to both. If a brand-new entity is first evoked by a referring expression that explicitly links it to another, already evoked, entity, that entity is said to be *brand-new anchored* (Example 4.11.d).

(4.11) a) [Poochie]$_{\textit{brand new unanchored}}$ is a nice little dog owned by my neighbours.
b) [He]$_{\textit{textually evoked}}$ is very friendly to strangers.
c) [You]$_{\textit{situationally evoked}}$ will hear more about the little bastard later on in the examples of [this thesis]$_{\textit{situationally evoked}}$.
d) Notice that [the standard of these examples]$_{\textit{brand new anchored}}$ is declining rapidly,
e) in particular that of [the example you are reading at the moment]$_{\textit{containing inferable}}$
f) In [good linguistic example writing tradition]$_{\textit{inferable}}$,
g) I will finally say something completely unfunny about [Kofi Annan]$_{\textit{unused}}$.
h) What a relief to be spared the usual U.S. president for a change!

Later, Prince (1992) reduced that detailed scheme to two binary taxonomies: discourse-old/new versus hearer-old/new. Hearer-old entities are unused or evoked, while discourse old entities are textually evoked. Discourse old/new corresponds to the distinction between first and subsequent mentions, which can be labelled reliably even by relatively untrained annotators. Lambrecht

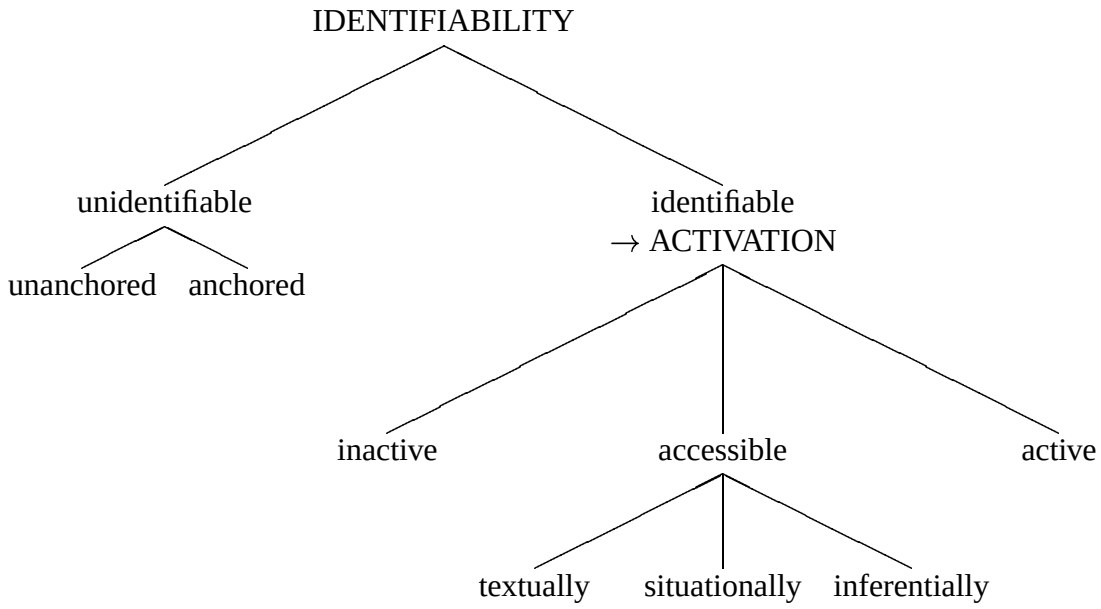


Figure 4.4. Taxonomy of givenness according to Lambrecht (1994, Diagram 3.25, page 109)

(1994) integrates Prince's account of assumed familiarity with cognitive considerations. He differentiates between *identifiability* and *activation* (c.f. Fig. 4.4). Whether a discourse entity can be identified depends on what the addressee already knows about it, or how familiar he is with it. How strongly activated it is, on the other hand, is a matter of consciousness and attentional state.

Lambrecht integrates identifiability and activation into the hierarchy given in Figure 4.4. This hierarchy is the basis for the largely source-based coding scheme of entity status that I propose in Section 5.2. Unidentifiable discourse entities correspond to Prince's brand-new ones. In contrast to Prince and to the source-based scheme from Section 5.2, Lambrecht proposes no further subcategorizations of inferential accessibility. The taxonomy also differentiates between three activation states: inactive, accessible, and active. These three states roughly correspond to Chafe's inactive, semiactive, and active (c.f. below Section 4.3.6).

An *active* discourse entity sits in the addressee's working memory; it has been mentioned very recently in the discourse. If an active entity does not get mentioned for a while, it ceases to be active and becomes *textually accessible*. However, a previous mention in the text is not he only way of making a discourse entity accessible; it can also be evoked by scripts. For example, when talking about planes, our cultural background immediately evokes the concept of a pilot in us. Lambrecht calls this type of accessibility *inferential accessibility*. The discourse entity can also be accessible because they are present in the situational discourse context. This is *situational accessibility*. For example, if you have a printed copy of this thesis lying on front of you, the paper on which the thesis is printed is accessible from the discourse situation. Finally, an inactive discourse entity can be identified by the addressee, but has not been mentioned for a while. In general, hierarchies that are based on familiarity, identifiability, or shared knowledge, for short, are more *procedural* than cognitive hierarchies. Instead of just characterising the cognitive state of a discourse entity, they specify where additional information about a new entity can be obtained. Take the following discourse:

	$C_b(U_i) = C_b(U_{i-1})$	$C_b(U_i) \neq C_b(U_{i-1})$
$C_b(U_i) = C_f(U_i)$	continue	smooth shift
$C_b(U_i) \neq C_f(U_i)$	retain	rough shift

Table 4.3. Types of transitions between utterances, modified after (Brennan et al. 1987)

(A plane) $_{PL}$ crashed into the ground.

(The pilot) $_{PI}$ managed to escape with a shock before it burst into flames.

Let's assume that the speaker is giving this example at a linguistics colloquium. In the first sentence, a new referent "a plane" (*PL*) is introduced; the expression is referential. The hearer constructs a mental representation for that plane and adds both the plane and the plane crash to his focus of consciousness. After the sentence has been uttered, *PL* is active. Planes usually have pilots, so that with the mention of a plane, the pilot of that plane (*PI*) would become semiactive. When the pilot is in fact mentioned in the second sentence, the referent of that expression is therefore uniquely identifiable, and the pronoun referring to the plane can be resolved because (*PL*) is active. But while the cognitive account merely states that (*PI*) has a certain status, the familiarity hierarchy tells us why: because it is part of the frame evoked when mentioning planes.

4.3.2 Centering Theory

Centering Theory (Grosz et al. 1995) is a theory of local coherence within a discourse segment. Discourse segments are assumed to be hierarchically structured according to (Grosz and Sidner 1986). The theory is based on tracking (direct or indirect) realisations of discourse entities. An entity is directly realised in an utterance if there is a referring expression that points to it. The criteria for indirect realisation largely depend on the underlying semantic theory.

Each utterance U_i in a discourse is associated with a list of *forward-looking centers*, C_f. The connection between U_i and the preceding U_{i-1} is established via the *backward-looking center* $C_b(U_i)$, a discourse entity that has been realised in both utterances. There is at most one (in most versions: exactly one) C_b per utterance U_i, and this C_b can only be chosen from $C_f(U_{i-1})$. The backward-looking center should be the highest-ranking element in $C_f(U_{i-1})$ mentioned in U_i. The elements on the C_f list are partially ordered: the more likely it is that a subject will be C_b of the following utterance, the higher its rank.

Transitions between utterances are classified according to two criteria: $C_b(U_i) = C_f(U_i)$ (whether the backward-looking center of an utterance is its preferred forward-looking center) and $C_b(U_i) = C_b(U_{i-1})$ (whether the backward-looking center is maintained across utterances). The commonly used types of transitions are summarised in Table 4.3 (Brennan, Friedman and Pollard 1987, Walker, Iida and Cote 1994). Further refinements such as costs on transition pairs and additional transition types are discussed in (Strube and Hahn 1999).

Grosz et al. (1995) do not claim that centering can explain for any stretch of discourse taken from a contiguous discourse segment whether it is locally coherent or not. Rather, they set out

to model constraints on referential cohesion on the basis of the two data structures defined above, the C_b and the ranked C_f-list. Extrapolating from (Grosz et al. 1995, Section 5), we can conclude that a fully worked out Centering Theory should

1. constrain the form in which discourse entities are realised in the utterances of a discourse segment, depending on whether they are the current C_b or were the old C_b, and depending on their rank on the C_f list. Currently, only one constraint on pronouns has been stipulated:

   Definition 4.3 (Rule 1 of Centering) *If any element of the list of forward-looking centers $C_f(U_{i-1})$ is realised by a pronoun in U_i then the C_b of U_i must be realised by a pronoun as well. (Grosz et al. 1995)*

2. constrain the transitions between utterances in a segment. This is captured by Rule 2:

   Definition 4.4 (Rule 2 of Centering) *An interpretation that yields a continue transition is preferred over one that yields a retain, and retain transitions are preferred over shifts.*

3. integrate syntactic, semantic, *and* pragmatic influences on local coherence

The actual constraints on the rank of the forward-looking centers are hotly debated. The classical implementation of (Brennan et al. 1987) uses mainly grammatical roles:

(4.12) SUBJECT > DIRECT OBJECT > INDIRECT OBJECT > COMPLEMENTS > ADJUNCTS

A number of researchers have proposed language-specific variations (see e.g. Walker et al. 1994, Turan 1998, Hoffman 1995, Di Eugenio 1998), For example, Walker et al. (1994), working on Japanese, made use of the fact that Japanese codes topicality with a special particle, *-wa*. Accordingly, they rank topics highest on the center list. Second comes the entity with which the speaker empathises, followed by the grammatical functions subject (postposition *ga*), object2 (postposition *o*), object (postposition *no*) and other arguments. The complete hierarchy is as follows:

(4.13) (GRAMMATICAL OR ZERO) TOPIC > EMPATHY > SUBJECT > OBJECT2 > OBJECT > OTHER

Other researchers have explored possible language-independent constraints on C_f ranking. Cote (1998) proposes to rank the elements of the C_f-list in terms of the underlying Lexical Conceptual Structure (LCS) of the sentence. LCS describes the semantics of a sentence in terms of a restricted set of semantic primitives, which inter alia allows to categorise verb arguments according to their thematic role. Since LCS provides a principled way of determining thematic roles, Cote's approach subsumes orderings based on thematic roles such as those of Turan (1995, 1998). Another avenue that has been explored is what has been termed "information structure". Strube and Hahn (1996), inspired by (Daneš 1974a), partition the elements of the C_f-list into a given element (the C_b), a thematic (the C_f) and other elements, which are more or less contextually bound. In their recent journal publication, however, (Strube and Hahn 1999) switch to an ordering based on Prince's (1981) familiarity hierarchy.

Although Centering Theory has been very influential, it can be criticized on numerous counts. Strube and Hahn (1999, page 339f.) list the following points. First it is not at all clear what counts as an utterance, hence, it is not clear how to process complex sentences; for possible solutions, see e.g. (Suri and McCoy 1994, Kameyama 1998). Complex, nested NPs present another problem (Walker and Prince 1996). The model is also fairly restricted in the range of anaphoric expressions that it covers. The model has barely been implemented computationally so far; and comparative large-scale evaluations are scarce. Even worse, the first such evaluation, reported by Poesio (2000), suggests theory-internal problems with Centering. They tested different formalizations of utterances, different methods of ranking the C_f list, and whether implicit realisations should be counted as well as explicit ones. They found a tradeoff between Rule 1 (c.f. Definition 4.3) and the constraint that each utterance may have only one C_b. As recent incremental models of pronoun resolution suggest, pronoun resolution might be less affected by discourse segment boundaries than many researchers assume Walker (2000) retains Centering as a model of local coherence, but suggests that it should also be applied across segment boundaries. She replaces the stack of focus spaces that Grosz and Sidner (1986) proposed by a more flexible cache model of attentional state. Strube (1998) is even more radical; he discards the notion of a C_b altogether. Instead, he proposes a list of potential antecedents ordered mainly according to their familiarity. That list, called S-list, consists of expressions mentioned in the current and in the preceding utterance. In order to compute his ordering, he uses both surface position and a tripartite familiarity scale based on (Prince 1981), but modified according to how accessible an entity is to the hearer: *old* entities are those discourse entities that the hearer already knows, be it from world knowledge or from the preceding discourse (Prince's "unused" and "old"), *new* entities are brand-new unanchored discourse entities, and all other discourse-new entities are classified as *mediated*, because access to them is either mediated by an explicit anchor or by a bridging inference.

4.3.3 Accessibility

Where Centering is restricted to local coherence, and, in its current state, mostly to predictions about pronouns, Mira Ariel's (1990) Accessibility Theory purports nothing less than to explain how the system of referring expression of a language is organized. For this purpose, she resorts to a functional cognitive explanation: the more phonetic and other linguistic information a referring expression contains, the more information is needed to access the corresponding discourse entity, and the more information needed to access a discourse entity, the less accessible it is. Ariel does not distinguish discrete degrees of accessibility. Instead, she orders all forms of referring expressions of a language on a scale of increasing accessibility. Table 4.4 reproduces that hierarchy for English. Ariel makes the strong prediction that when it comes to choosing the form of a referring expressions, it does not matter which source the discourse entity comes from. All that matters is how easily accessible it is, be the source the co-text, the physical context, or world knowledge. She refines that point with respect to deixis in (Ariel 1998). The only source-based distinction that should affect the distribution of referring expressions should be whether the required discourse entity already exists in the discourse model or whether it still needs to be integrated. This prediction is partially confirmed by the analysis of radio news data presented in Section 6.3.

For Ariel, accessibility is a compound variable, which is determined by four more specific

Accessibility	**Form**		**Example**
Low	*Full name*	+ modifier	Mr Gerhard Schröder,
			the German chancellor, said
		(known referent)	Gerhard Schröder said
	Definite Description	long	the chancellor of Germany said
		short	the chancellor said
	Last name	Schröder said	
	First name	Gerhard said	
Mid	*Demonstrative*	distal + modifier	that elegant man said
		proximal + modifier	this elegant man said
		distal (+ NP)	that guy said
		proximal (+ NP)	this guy said
	Pronoun	stressed + gesture	HE (pointing) said
		stressed	HE said
High		unstressed	he said
		cliticized	said 'e
	gaps, agreement, reflexives		said to himself

Table 4.4. Accessibility Marking Scale after (Ariel 1990, page 73). The scale is continuous. Accessibility increases from top to bottom.

factors:

- *distance* to last mention
- the number of competing antecedents (*competition*)
- *salience*, which she operationalises as topicality, and
- the presence or rather absence of any discourse structural boundaries between antecedent and anaphor (*unity*).

Ariel does not claim the list to be exhaustive. We will meet the first two factors again in Chapter 7, where we will see that they are exceptionally robust (*competition*) and powerful (*distance*) predictors of pronominalization. While these two factors are relatively easy to measure on corpora, the second pair is not. *Salience* is in itself a notoriously muddy notion, which has been taken by other linguists to cover much of that what Ariel would term accessibility. *Unity* summarises in Ariel's definition (1990, page 29) all aspects of discourse structure that might potentially become relevant, such as point of view or paragraphs.

In her own corpus analyses, Ariel tends to reduce the complex variable of Accessibility to that of its components which is easiest to determine from corpora: distance to last mention (c.f. also Section 5.5). For almost each pair of adjacent forms on the hierarchy, she shows that they differ significantly with respect to the average distance to their antecedent. Strictly speaking, such analyses only show that distance to last mention is related to the form of referring expressions, but say nothing about the compound variable of accessibility.

Four aspects are remarkable about Ariel's work. Firstly, she strives to account for a very large range of phenomena. She even claims that Binding Theory can be annihilated if reflexives

State	Definition
type identifiable	hearer knows which type of object is referred to
referential	hearer needs a representation for a specific object
uniquely identifiable	hearer can identify specific object on the basis of the referring expression
familiar	hearer has a mental representation of the object referred to in memory
activated	object representation is in hearer's short-term memory
in focus	object representation in current center of attention

Table 4.5. The Givenness Hierarchy (Gundel et al. 1993)

and reciprocals were to be analysed with a more fine-grained accessibility scale (see (Ariel 1994) and (Ariel 1990, Part II) for details). Second, she substantiates most of her claims by corpus analyses. Third, by working on a Germanic, Indo-European language, English and a Semitic one, Hebrew, she adds an exciting typological dimension to her work. Finally, she applies her results to the sociological analysis of texts. She found that indeed, women tend to be referred to by forms that ranked higher on the accessibility scale, such as the first name, while men tended to be referred to by Low Accessibility expressions.

It cannot be doubted that Ariel uncovered extremely interesting patterns of gradation. However, there are several problems with the model as it stands. The central variable of accessibility is not specified very precisely, the contributions of the various factors are not weighted, and the interaction between those factors is not very clear. This is not a problem that can be solved by any short-term research programme because Ariel has defined her accessibility to be nearly all-encompassing. Furthermore, the interaction between accessibility and the choice of lexical forms is not quite clear. But since for Ariel, Accessibility Theory is firmly embedded into Relevance Theory (Sperber and Wilson 1995) as the means for accessing discourse entities and the corresponding contexts for further evaluation, we may expect that this is a problem whose solution will be taken care of by that theory.

4.3.4 The Givenness Hierarchy

Gundel et al. (1993) present an implicational hierarchy of cognitive states which they call the Givenness Hierarchy (Table 4.5). This hierarchy is restricted to discourse entities. It describes potential cognitive states of such referents in the mind of the hearer. If a speaker assumes that a discourse entity has a certain status, then this constrains the surface linguistic forms she can use to refer to that entity: If she is cooperative, she should only use a form that meets *at least* that status. In other words, unless most other proposals we have discussed in this Section 4.3, the Givenness Hierarchy is implicational. While all noun phrases in a discourse are at least type identifiable (corresponding to the lowest level on the hierarchy), only very few are in focus (at

the highest level on the hierarchy) at any given time in any given discourse. But that an entity is "in focus" does not mean that it has to be referred to by a pronoun. In principle, we are free to exchange that pronoun for any form that occurs on a lower level of the hierarchy. For English, Russian and Spanish, Figure 4.5 shows for a number of determiners which status a discourse entity needs to have before they can be used in a referring expression that specifies that entity.

(4.14) I couldn't sleep last night.
a) [A dog next door]$_{\textit{type identifiable}}$ kept me awake. I don't know which one from the damned brood my neighbours keep that was, I don't even know if the stupid beasts took turns yelping at the moon. Anyway, whoever that was, [he]$_{\textit{type identifiable}}$ had healthy lungs.
b) [This dog next door]$_{\textit{referential}}$ kept me awake. It was howling so pitifully that I had to get up and check what was going on.
c) [The little Yorkshire terrier of my neighbours]$_{\textit{uniquely identifiable}}$ kept me awake last night.
d) [Poochie]$_{\textit{familiar}}$ kept me awake all night again with his howling.
e) [That]$_{\textit{activated}}$ kept me awake. (Immediately before the sentence is pronounced, Poochie lets out a pronounced howl.)
f) Poochie has these howling attacks from time to time, you know. [He]$_{\textit{in focus}}$ kept me awake.

The six states and their definitions are summarised in Table 4.5. All noun phrases are at least type identifiable. *Type identifiable* noun phrases do not refer (Example 4.14.a). The definition of type identifiability leads to an interesting problem. If generic noun phrases do not refer, then they are type identifiable, no matter whether they are realised as a pronoun or as a full NP. In long stretches of discourse about kinds as in Example 2.8, page 9, the natural kind is bound to be pronominalised some of the time. But the theory assumes, as we will see below, that pronouns should only be used for entities that are at least activated. The resulting problem can be solved if we restrict ourselves to expressions that really refer, but this solution excludes data that less semantically conscious theories such as Ariel's Accessibility Theory will have no problems describing.

A *referential* expression indicates that the speaker has a specific referent in mind (Example 4.14.b). A particular case I encountered in my data appears to be on the borderline between the two categories. When news writers use a phrase such as "a spokesman from the White House", the name of the spokesman is irrelevant, the person is mainly referred to in order to attribute the information correctly. When there is a real event, such as a press conference, behind such a noun phrase, then there was a spokesman there who imparted the information, and the expression can be taken to refer to that particular person. But if we just focus on the attribution function that this NP fulfils, we can also label this NP as merely type identifiable. The next level of specificity is *uniquely identifiable*. The distinction between this and the referential level is subtle: whereas the complete sentence is needed in order to determine the referent of an expression that is merely referential, a uniqely identifiable referring expression provides all necessary information in the noun phrase itself (Example 4.14.c).

Familiar discourse entities are known to the addressee, either because they are part of world knowledge (the standard example is the current American president, who at the time of writing

was William Jefferson Clinton), or because they have been mentioned in the preceding co-text. The next category, *activated*, is defined in almost purely cognitive terms. Activated discourse entities are in current short-term memory; they would correspond to the explicit focus set of Sanford and Garrod (1998). But discourse entities need not be in short-term memory to be activated; immediate physical co-presence is also sufficient. The final category is *in focus*. The following quote illustrates that "in focus" means "in the current focus of attention" here.

> The entities in focus at a given point in discourse will be that partially-ordered subset of activated entities which are likely to be continued as topics of subsequent utterances. Thus, entities in focus generally include at least the topic of the preceding utterance, as well as any still-relevant higher-order topics. (Gundel et al. 1993, page 279)

Gundel (1988) defines the topic of a sentence to be that which the sentence is about. Note that Gundel, Hedberg, and Zacharski say nothing about exactly how likely is "likely". They also leave the door wide open for suddenly resurging discourse topics, which happen to get mentioned again after a while during which they have stayed respectfully in the background. Entities can be brought into focus by syntactic means such as topicalisation or by prosodic means. But linguistic means do not determine completely what will be focussed. Gundel et al. (1993) cite the example of a pronoun that refers to an adjunct in the preceding clause. They argue that if the adjunct is somehow "salient" in the context, the pronominal reference is licensed, because the entity that the adjunct specifies has obviously been brought into focus by some mechanism, as in Example 4.15 (Gundel et al. 1993, Example 11, page 280). If the adjunct is not pragmatically salient, the pronominal reference would be out, as in Example 4.16 (Gundel et al. 1993, Example 10, page 280).

(4.15) However, the government of Barbados is looking for a project manager for [a large wind energy project]$_{WP}$.
I'm going to see the man in charge of [it]$_{WP}$ next week.

(4.16) Sears delivered new siding to my neighbors with [the bull mastiff]$_{BM}$.
[It's]$_{BM}$ the same dog that bit Mary Ben last summer.
Anyway, this siding is real hideous and . . .

Although these observations capture interesting patterns of language in use, I do not see how they could be translated into an annotation manual without avoiding circular instructions, such as "if there is a pronoun, try every trick of interpretation that appears somehow reasonable to you to justify that it's in focus". Nowhere does it surface more clearly than in these instructions for distinguishing "activated" from "in focus" items what the Givenness Hierarchy really is: a set of common-sense conventions for protocolling the intuitions an addressee has about the accessibility of discourse entities.

Treated this way, the Givenness Hierarchy is an extremely valuable research tool if we aim to investigate how different people with different background react to and understand the same text. What is more, Gundel et al. (1993) have shown that the categories of the Givenness Hierarchy can be applied profitably to analysing data from typologically very different languages: Chinese, Japanese, English, Russian, and Spanish. However, it does not tell us very much about

	in focus >	activ. >	fam. >	uniq. id. >	refer. >	type id.
English	it	HE, this, that, this N	that N	the N	indef. this N	a N
Russian	zero, on ('he')	ON, eto ('this'), to ('that')	eto N, to N	zero + N		
Spanish	zero, el ('he')	EL, este (proximal), ese (medial), este (distal)	ese N (medial), este N (distal)	el N ('the N')	un N; zero + N	

Figure 4.5. Highest required status of specified discourse entity for determiners in English, Russian, and Spanish. Capitals: stressed pronoun. Zero: no surface referring expression. Source: (Gundel et al. 1993, Table 1, page 284)

how discourse entities are managed. The reason is that, as Chafe (1994) has noted, it conflates two related dimensions, identifiability and activation/accessibility. Both dimensions are important, that is why it will turn out to perform so well in the analyses of Chapter 6, but conceptually, it might be better to separate them.

4.3.5 Grammar as Mental Processing Instructions: Givón

Talmy Givón has made two important contributions to the study of how discourse entities are managed: He proposed empirical measures of the distribution of referring expressions, and he developed a theory of grammar as mental processing instructions which states that the referring expressions function as instructions for constructing and retrieving discourse entities. He has developed and refined both contributions over the years (Givón 1983a, Givón 1992, Givón 1995b, Givón 1995a). This section is based on two recent statements of his position, (Givón 1992, Givón 1995a). We will first discuss the measures, then the conclusions he has drawn from his corpus data about the mental instructions of reference processing.

Corpus-Based Measures: The quantitative measures proposed by Givón (1983a) are easy to annotate, and thus do not require complicated inferences about a speaker's intentions and a hearer's consciousness. I will focus on the two most popular of his measures, Referential Distance (RD) and Topic Persistence (TP) here. Both measures are based on co-specification sequences.

Referential Distance is distance to last mention with a twist: All distances greater than 20 clauses are mapped to the value 20, which was fixed arbitrarily; first mentions are also assigned this value. The resulting measure is called Referential Distance (RD). Besides referential distance, which covers the anaphoric dimension of co-specification sequences, Givón has also developed a count for the cataphoric dimension: Topic Persistence (TP). This measure counts how often an entity recurs in the following stretch of discourse. The more often it recurs, the more central to the discourse segment it is. Since no discourse segmentation is presumed, Givón posits an arbitrary limit of 10 clauses after the current clause. From the research reported by Givón, it appears, however, that TP values cluster into two large groups: those between 0 and

2, and those larger than two. Since TP depends to a large extent on the position of a referring expressions in a discourse segment, it is neither a direct measure of the topicality of an entity in a discourse segment, nor does it tell us when exactly the entity is mentioned next.

Both measures have their problems. Let us discuss RD first, because it is the most widely known measure. Firstly, it lumps first mentions together with mentions of entities that have not been accessed for a while. As Chafe (1994) argues, discourse entity management is multi-faceted, and one of these facets is clearly how new entities are introduced into the discourse. Introducing a new entity into the discourse and accessing an existing entity are different operations. True, if the addressee forgets a discourse entity that has last been mentioned many sentences ago, then it is effectively re-introduced when it reappears in the discourse. But whether an addressee forgets an entity depends on both his memory and on the role that the entity has played in the discourse so far. When he posits a cut-off value of 20, Givón effectively averages over all these influences. I believe that this is methodologically problematic. Either we make our measures as independent of the addressee as possible, then we cannot motivate the threshold any more, or we strive for a cognitively realistic model, which means that the RD measure itself becomes questionable and needs to be supplemented by a measure that characterises the role that an entity has played in the discourse so far. In the corpus study of co-specification sequences reported in Chapter 5.4, I have dealt with this issue by assigning a special status to first mentions. For Chapter 7, a special *categorical* variable, DIST, was defined that codes both distinctions.

At first sight, TP appears as arbitrary as RD. But there is more to that measure than one might think. Once we can derive our counts automatically from co-specification sequences, the fixed window length is not as important any more as it was at a time where counts had to be done manually. Thus, one of the main points of contention, the fixed window size, can be eliminated. Moreover, TP has an interesting mathematical property: it can only change by -1, 0, or 1 as we move through a text clause by clause; therefore, it is as smooth as discrete measures can get.

Processing Referring Expressions: Givón himself vividly denies that any of his measures have any direct cognitive correlate. Nevertheless, the quantitative data they have allowed him to collect, together with his quest for giving linguistics a solid biological and anthropological basis, have led Givón to a very detailed cognitive theory of how referring expressions are processed. Kintsch (1995) has argued that this theory fits very well with his Construction/Integration (C/I) model of discourse comprehension, arguably the most influential in modern psycholinguistics.

Givón assumes that grammar has evolved to be a fast, robust, rough mechanism that helps addressees to construct both locally and globally coherent discourses from the input that communicators give them. But *grammar-cued* coherence, as Givón calls it, is always secondary to *vocabulary-cued* coherence. By this term, Givón means the knowledge-based processes which are so central to both a Mental Models approach and SMF theory, but which are also part and parcel of the C/I model, and which, as Garnham et al. (1992) point out, are even stronger than grammatical gender cues in pronoun resolution. Vocabulary-cued processes are slow, but only they provide the necessary information for grounding new information in world knowledge.

Givón envisions discourse to be represented in the mind as a network of nodes, again in accordance with the C/I model. The information from incoming clauses is filed under special *topic chain nodes*. Topic chain nodes are in turn grouped under paragraph nodes. If we assume

that topic chain nodes correspond to discourse segments, and if we further assume that paragraph nodes correspond to segments at a higher hierarchical level, etc., we can directly derive a notion of discourse structure from these terms. Clauses for Givón are both major and minor clauses, not the large Major Clause Unit that we will introduce later in Chapter 5, Definition 5.7. Clauses are organized so that they impart at most one new item of information based on at least one item of already known information.

There are two main classes of operations: *attentional activation* operations that open and close topic nodes and *search and retrieval* operations that look for the grounding instance of a referent.

Topic nodes stand for a stretch of discourse with a coherent topic. In each clause, there is a *topical referent* (= topical discourse entity) which determines the node under which the information from that clause is filed. This referent may also be "realised" as a zero pronoun. As long as new information about a topic keeps flowing in, that topic is said to be *active*. When the topic is changed, the old topic needs to be closed and another one opened. Grammar cues topics via syntactic function: primary topics tend to surface as subjects, secondary ones as objects. Since topics are usually kept, not changed, he claims that "[z]ero lexical marking is . . . the default choice in the grammar of referential coherence". Whenever that principle is broken for a topical referent, it is likely that a new topic chain node needs to be activated. Only one topic node can be activated at a time.

Any referent that occurs in the discourse must be *grounded* in a phrase that contains at least one lexical morpheme which can make the appropriate connection to world knowledge.[7] Givón posits the following principle of grounding.

> A node—and thus the referent-label that activates it for text-storage—must be grounded before it can be activated for text-storage.
> a. A new (indefinite) referent is grounded only to its current text location in the episodic structure still under construction.
> b. An old (definite) referent is attached to its current location in the episodic text structure; but it must also be grounded to some other location in some pre-existing mental structure.
>
> (Givón 1995a, page 102)

This citation shows nicely the pros and cons of Givón's framework. On one hand, he presents a detailed model, couched in terms compatible with modern psycholinguistic theory. On the other hand, the processing instructions he comes up with on the basis of his corpus work are sweeping. For example, judging from the texts I have analysed so far in my comparatively short life as a linguist, new referents do *not* tend to be indefinite, at least not if they are to become important later on in the text. Introducing new discourse entities with the indefinite is a pattern one often finds in narratives, but in genres such as radio news (c.f. Chapter 6) this is very unusual. In fact, in the radio news texts the distribution of definites is such that it neither favors first-mention nor anaphoric uses. I would not hold that pattern against the radio news writers; rather, it appears to be popular in texts with a high informational density. The same skew can be found in the BROWN-COSPEC data (c.f. Table 7.7).

[7] Givón restricts himself to nouns in (Givón 1995a), but since antecedents of pronouns can also be verb phrases or longer stretches of discourse, I have chosen a more general formulation here.

In sum, Givón's work is a highly interesting approach on a solid empirical and theoretical basis. From the perspective of discourse entity management, he focuses mainly on implementing access mechanisms. Initialisation is given somewhat short shrift, at least in (Givón 1992, Givón 1995a). Givón emphasises that new entities are connected to the discourse network as soon as possible; he also states that the more connections, the easier an entity is to access. On the other hand, his model also covers aspects of structural entity status, because his network model is essentially a connectionist model of discourse structure. In contrast to e.g. Ariel's Accessibility Theory Givón's model extends easily to other linguistic markers of cohesion. It can also be implemented computationally: there exist implementations of its cognitive base model, the C/I theory (Kintsch 1988, Kintsch 1995), and Givón (1995a) has formulated an explicit set of major grammar-cued operations for establishing referential coherence that might provide a starting point for more thorough explorations.

4.3.6 Activation and Consciousness: Chafe

Finally, we come to an approach that is maybe closest in spirit to the perspective on communication advocated by Gerold Ungeheuer, the work of Wallace Chafe. Like Givón, Chafe works from empirical data, and like Givón, he is interested in the cognitive foundations of language. The present summary of Chafe's position is based mostly on (Chafe 1994), where he summarises his positions and makes some more general methodological points. All following citations refer to that work.

Chafe's theory of discourse comprehension is based on consciousness. To him, consciousness "is an active focusing on a small part of the conscious being's self-centered model of the surrounding world." (page 28). That self-centered model could be the Personal Experience Theory (PET) of each person that Gerold Ungeheuer postulated (c.f. Appendix D). What is in the focus of attention is *active*. Around that active area, we have a block of *semiactive* information (or peripheral consciousness) which we have met earlier on page 52 under the heading *discourse topic*. Since the focus of attention continuously changes, information keeps drifting into and out of focal and peripheral consciousness. This continuous flow is the *information flow* which is such a central metaphor in Chafe's work. Chafe interprets the term "information" very widely to mean events, states, and concepts. Thus, he can also talk about the givenness of a verb or an adjective.

On this conceptual basis, Chafe describes the givenness of a piece of information using three *activation states*, active, semi-active and inactive. *Active* information is in the "focus of consciousness" (Chafe 1987, page 25). *Semi-active* information are concepts which a person is aware of, but not focusing on at the moment, while *inactive* information has to be fetched from long-term memory when needed. Information can be semiactive for three reasons:

- either it has been mentioned earlier in the discourse,
- or it is associated with information that is or was active, and has been activated in long-term memory as a result,
- or it is accessible via physical co-presence in the communication situation (c.f. Clark and Marshall 1981)

The activation state an entity determines how much it costs to reactivate. The higher the cost, the more help the hearer needs (and hopefully also gets). Although this taxonomy is useful for describing what goes on in the minds of speaker and hearer during discourse, its operationalisation is quite difficult (Schütze-Coburn 1994)—no wonder, because Chafe's activation states protocol what goes on in the mind of a hearer who has to process a new chunk of discourse, and it also depends on the PET of that hearer what remains in his consciousness, what he decides to attend to, and what drifts away quickly.

Like Lambrecht (1994), who heavily built on Chafe's work, Chafe distinguishes between the identifiability of a discourse entity and its activation status. Identifiability means that the hearer can connect the referring expression to an existing *discourse entity*, as becomes clear from Chafe's discussion of generic referents on page 102f. True, the more highly activated an entity is, the easier it is to identify, but all that counts in the end is that the hearer knows what the speaker is talking about—in particular when that entity is new to the discourse. Therefore, Chafe insists that identifiability and activation should be kept strictly separate. Whether a discourse entity is easy to identify depends among other factors on whether it belongs to the common ground shared by speakers and hearers, and how salient it is in the current conversation. Chafe defines (contextual) salience here as the degree to which a discourse entity "stands out" from other discourse entities that might be categorized in the same way. But identifiability is not important for all nominal expressions. In particular when they occur in idioms or near-idiomatic collocations, it is not important to identify the referent of the noun phrase itself, but to identify the event or state that is being reported by the idiom.

Speakers present information to their hearers in handy chunks, *intonation units*. These units are usually small, so that they easily fit into working memory. They are subject to what Chafe calls the "one new idea" constraint: there should be at most one new piece of information in an intonation unit, and at least one given one. Subjects have a special status in this respect: they mark starting points for processing the message. Ideally, starting points should be given and active. This led Chafe to formulate the *Light Subject* constraint: Subjects carry a light information load. "Light" does not equal "given", however. Although most subjects in conversation are indeed both given and active, some of them are merely accessible (semiactive). If a subject expresses new information, for example if it introduces a brand-new discourse entity, then that information is trivial and the new discourse entity will not be important to the discourse. Note again that the criteria Chafe posits are based heavily on interpretation and introspection. Although people appear to agree quite well on which discourse entities are important in a text and which are not (Wright and Givón 1987), it is certainly possible to devise any hard and fast criteria. The only approximation to such a measure is frequency of measure.

Chafe's theory is perhaps the most "communicative" in the sense of (Ungeheuer 1967/1972a) that we have surveyed so far. Since he intends to explain how the flow of information in discourse is signalled by linguistic means, he avoids the temptation of one-dimensionality to which Ariel succumbed so eagerly. In fact, Chafe views her work as an interesting exploration of what he has termed "activation cost", but clearly sees its limits. Since Chafe resolutely takes a communicative stance, annotating a discourse with his categories requires much introspection and interpretation. This is not necessarily a bad thing; in fact, we need such research in order to give meaning to quantitative and experimental results. On the other hand, a linguist that analyses a text, as Brown (1995) has argued so incisively, will never be able to completely re-construct the perspective of those who produced them. Even worse,

a linguist needs to approach a discourse with a heightened attention to detail, and an eye for minute connections, that may easily lead her to interpretations that were not intended by the addressees.

Finally, let us summarise Chafe's approach in terms of how discourse entities are managed. He covers *initialisation* only indirectly, because his focus is not on the point at which a discourse entity becomes part of a discourse model, but on the point at which it is introduced into the peripheral consciousness. For our purposes, we can say that if an entity is active or textually accessible, it is in the discourse model—although it is tempting to extend the discourse model to what is only semiactive, but has never been active before. This temptation can only be resisted if we assume that it makes sense to distinguish between world knowledge that is more or less relevant to the current discourse and a trace of the past discourse co-text. Chafe's semiactive discourse-new referents provide a nice way of modelling some bridging inferences.

Like Givón, Chafe is more concerned with describing access than initialisation. Chafe describes the contribution of consciousness to access via activation states, and the contribution of discourse context, community membership, and physical co-presence via identifiability. He addresses the question of update in his discussion of how intonation units are organised, anchoring new information to old.

4.4 Summary

If we want to model how discourse entities are managed computationally, we need to cover three main functions: initialisation, access, and update. The update function is perhaps the most complex one, because that has to take into account current beliefs, old beliefs, it has to do away with old beliefs, if necessary. The access function must manage conflicting beliefs, different access routes, pragmatic innuendo, lexical semantic detours, and pronominal highways.

In constructionist psycholinguistic models of discourse comprehension, these three functions are taken care of by a more general construct, the situation model or Mental Model. To access a discourse entity means to access an important structural component of the model, and to initialise a discourse entity means to either add a new component to the model, or to re-focus the model so that a mesh of properties and relations suddenly appears as a unit that can be referred back to later in the discourse. Our brief discussion of first mentions has shown that it is difficult to determine the precise time at which a relatively random mesh merges into a relatively fixed unit. A possible solution could be that after each utterance, a number of potential discourse entities are available to both speaker and hearer. Entities that were specified by noun phrases and that play an important part in the current discourse model are available longer than entities that would correspond to VPs or parts of the discourse. When an entity is successfully referred back to anaphorically, it has been grounded and becomes available as a bona fide "conceptual coat hook" (Woods / Webber).

Although many theorists attempt to model how discourse entities are managed, I have found no approach that is both cognitively plausible and easy to operationalise. However, a few researchers come close, Givón (1995a) with his resolutely functionalist cognitive model, and Chafe (1994) with his resolutely functionalist communicative model. The problem with Chafe's work is that his categories are very difficult to annotate. They require a good deal of interpretation and introspection. Givón's measures, on the other hand, are very easy to measure. They

can be derived automatically from co-specification sequences, which are, as I will argue in the next chapter and in particular in Chapter 5.4, about the only aspect of entity status that can be annotated quickly and reliably on large data sets. Since Prince's and Lambrecht's taxonomies need relatively detailed hearer models in order to be applied successfully, they will be explored further in Chapter 6 on texts from a very specific genre with a relatively well-researched, albeit complex, communication configuration.

5 Entity Status in Corpora

This chapter is an interlude between the theoretical part of the thesis (Chapters 2–4) and the empirical part (Chapters 6–7). It is dedicated to questions of methodology. The basic question I pose here is: How should corpora be annotated for studying linguistic correlates of entity status?

The section is structured as follows: First I critically review previous corpus-based studies of entity status (Section 5.1) and make some general remarks about the corpus-based testing of linguistic hypotheses. Then, in Section 5.2, I define the annotation scheme that I used in the annotation of the radio news texts. In Section 5.3, I discuss a quantitative measure of entity status: distance from last mention. Distance to last mention opens up exciting ways of statistical analysis. In Section 5.4, one of them is explored further: modelling co-specification sequences by stochastic processes. Finally, in Section 5.5, I critically evaluate the usefulness of measures such as distance from last mention for research on entity status.

5.1 Corpus-Based Research on Entity Status

It would lead too far afield to survey all large-scale corpus-based studies of anaphora here. Therefore in Section 5.1.1 I present an overview of commonly used methodologies and discuss one cross-genre study in detail, that of Biber (1992). Then, in Section 5.1.2, I focus on an aspect that is particularly important for the following empirical chapters, annotation schemes for sequences of antecedents and anaphors. Finally, Section 5.1.3 presents some conclusions.

5.1.1 Corpus-Based Studies: Some Examples

Many classic studies are corpus-based: Chafe's (1980) Pear Stories corpus, which has been the basis for much subsequent research, the corpus analyses documented in (Givón 1983c), Fraurud's (1990) study of non-anaphoric definites, Gundel et al.'s (1993) Givenness Hierarchy, and Ariel's (1990) Accessibility Theory. Their results have already been discussed extensively in Chapter 4; now, I will summarise their methodology.

A very common method is to take arbitrary texts, often from magazines, journals, and novels, and to analyse them. This appears to have been the procedure followed for many of the examples in (Ariel 1990). Brown (1983a) bases her study of Topic Continuity in written English entirely on the novel "Dr. No" by Ian Fleming, and Givón (1983b) uses a spoken monologue by a man from New Mexico. Although this procedure permits interesting qualitative insights, the results are definitely not representative.

Some researchers who want to make more general claims take care to cover several genres.

But they do not always specify the distribution of their samples across genres or domains, and sometimes even do not specify their sources. A good example for this is (Gundel et al. 1993). For studying patterns of co-specification in speech, more and more researchers record their own data, task-oriented monologues (Chafe 1980, Nakatani 1997) or dialogues (Hockey 1998).

In all studies for which the researchers have to collect the data themselves, the amount of data and the depth of analysis are severely restricted by the time that they can spend on gathering and encoding their corpus. If, on the other hand, the corpora you choose to work on are already annotated, e.g. with a syntactic parse, you can analyse more data more thoroughly and efficiently. This is the strategy that was followed by e.g. (Strube and Wolters 2000) for written language and (Francis, Gregory and Michaelis 1998) for speech. Apart from saving the analysts work, such a procedure has another crucial advantage: Working on publicly available, standard corpora makes the results more easy to replicate.

A number of corpora have been created as training data for anaphor resolution algorithms. The MUC corpus consists of hand-annotated newswire texts which were annotated for the Message Understanding Conference competitions using the specially designed Message Understanding Conference Coreference Scheme (MUCCS) was designed. One of the largest efforts is certainly the Lancaster Anaphoric Treebank, a large body of texts from American newspapers which was marked up with textual cohesion relations following (Halliday and Hasan 1976). The annotation effort is documented in (Fligelstone 1992, Garside, Fligelstone and Botley 1997). The corpus was built in order to serve as training data for anaphora resolution algorithms. da Rocha (1997) labelled parts of the London-Lund corpus and a corpus of Brazilian Portugese dialogues he collected himself with rich informations about anaphor-antecedent/sponsor sequences. In particular, he coded for each anaphor-sponsor pair the resolution strategy that needed to be applied in order to find the sponsor. He presents the results in great quantitative detail in what he calls his *antecedent likelihood theory* of anaphor resolution (da Rocha 1998). This theory largely consists of a structured summary of the patterns in his data, arranged in a decision-tree format.

Most corpus-based studies focus more on the linguistic patterns in their data and less on automatic induction of resolution or generation algorithms. From the large number of such papers, I will report on only one, which is particularly pertinent to the cross-genre research to be presented in Chapter 7, the study of Biber (1992). Biber looked at the distribution of referring expressions across genres in the London-Lund corpus of spoken British English (Svartvik 1990), and the Lancaster-Oslo-Bergen (LOB, Johansson, Atwell, Garside and Leech 1986) corpus of written British English. The genres of the LOB corpus are modelled on those of the Brown corpus which is used extensively in this thesis (c.f. Appendix C). Biber's general approach is to identify a set of easy-to-compute linguistic indicators, compute their frequency in all texts in his sample, and then run a factor analysis on the result in order to discover groups of texts that are obscured by the categories assigned by analysts (Biber 1988). In the 1992 study, he used the same method, but this time, he concentrated on features defined in terms of co-specification sequences and anaphoric expressions. Biber uses the term "referential chains" in his work.

He chose 58 texts from nine genres (Brown categories in brackets): press reportage (CA), legal documents (CH), humanities academic prose, technical academic prose (both category CJ), general fiction (CK), face-to-face conversation, sports broadcasts, spontaneous parliamentary speeches, and sermons. From each of these texts, he analyzed the first two hundred words. In

contrast, for BROWN-COSPEC, we used the full texts as contained in the corpus, but limited the number of texts to twelve. In the end, using fewer, but longer, texts should give a clearer picture of the distribution of referring expression than going for many brief excerpts from the beginning of texts which are themselves frequently excerpts from the middle of a longer piece of discourse Altenberg (1992).

Biber marked up the texts as follows. In a first pass, he automatically classified all nouns as referring and all first and second person pronouns as exophoric. He then established links between all nouns with the same lexical form, and computed the position of anaphors and antecedents (in prepositional phrase, in relative clause, in other dependent clauses, in major clause). In a third pass, he resolved all pronouns by hand and linked nouns that were not repetitions of their antecedents to the proper antecedents. Although this procedure is very simple and effective, it completely fails to take into account that the most basic unit in co-specification sequences are referring *expressions*. He also marks anaphor/sponsor pairs where the anaphor repeats the sponsor verbatim, but clearly does not specify the same discourse entity (Altenberg 1992). Since no parser or chunker was available to Biber, he had to restrict himself to preprocessing steps that are easy to implement on the basis of a tagged corpus.

He then extracted a number of features that described the frequency of different types of referring expressions and of types of co-specification sequences. These features were combined with the scores of each text on the five dimensions of genre as defined by Biber (1988). These features are summarised and commented in Table 5.1. A factor analysis yielded four referential dimensions:

Involved Referential Strategies: involved production (Biber's genre dimension 1), exophoric and discourse deictic/cataphoric pronouns, long co-specification sequences, higher average distances, few lexical repetitions.

"Named" Referential Strategies: more lexical repetitions, more sequences, higher average distances, less narrative focus (Biber's genre dimension 2)

Expository versus Narrative Strategies: more explicit reference (Biber's genre dimension 3), abstract style (dimension 5), overt persuasion (dimension 4), less anaphoric pronouns and narrative focus (dimension 2)

Referential Density: more new discourse entities, more entities mentioned only once, higher average distance, smaller average chain length, less overt persuasion (dimension 4)

Looking at the mean scores of each genre on these dimensions, we find that academic prose uses many "named" referential strategies and is very densely populated by referents, while general fiction uses these "named" strategies least often and also scores on the narrative end of the "expository vs. narrative" dimension. Spot news (very roughly) patterns with academic prose.

5.1.2 The Question of Annotation

In computational linguistics, most modern annotation schemes are based on SGML, the Standard Generalised Markup Language (Goldfarb 1990), or XML (eXtended Markup Language), a subset of SGML. In this section, I discuss two schemes for the annotation of co-specification sequences in more detail, the MUC scheme (MUCCS, Hirschman and Chinchor 1997) and the

1. total number of referential chains (roughly corresponding to co-specification sequences)
2. number of discourse entities that are only evoked once ("deadend")
3. average length of referential chains
4. average distance between two mentions
5. maximum distance (largest average distance in text)
6. first mentions
7. nouns that are repeated in the text (for Halliday and Hasan (1976), that would be part of lexical cohesion)
8. anaphoric pronouns (noun or pronoun antecedent in text)
9. exophoric pronouns (first/second person)
10. other pronouns (labelled by Biber as 'vague')

Table 5.1. The ten referential features (from an original total of 24) chosen for the final factor analysis.

MATE scheme (Poesio 2000). In that context, I will also discuss problems in annotating bridging inferences. For reasons of space, I will not discuss some of the alternative schemes that have been devised, such as those of Fligelstone (1992), Botley (1996), or da Rocha (1998). Instead, I concentrate on those schemes that have influenced my own work most.

SGML I: MUCCS. The most widespread SGML-based scheme is arguably that which was devised for the Coreference Task of the Message Understanding Conferences (MUC). The task is roughly specified as follows: Given a newswire text whose structure (headline, body, etc.) has already been labelled, find all referring expressions and the co-specification sequences that hold between them. Systems need not detect complete referring expressions; it is sufficient if they find the correct nominal heads. I will not go into the details of the scoring scheme (Vilain, Burger, Aberdeen, Connolly and Hirschman 1995) here. Instead, let us focus on the coding scheme. It was designed with a simple premise in mind: What human annotators cannot annotate reliably, machines cannot learn. Therefore, the designers took great care to ensure that the guidelines were so concise that annotators differed on as few decisions as possible, and the scheme has been revised several times. The current version is (Hirschman and Chinchor 1997), written for MUC-7.

Co-specification sequences are coded as sequences of *markables* in the text. Two markables co-specify if they access the same discourse entity. This means that MUCCS only allows for *identity* relations between referring expressions and their sponsors or antecedents in the text. The first extensional reference in a sequence is called the *grounding* instance. It connects the sequence to an individual in the world.

Attribute	Description
`id`	a unique identifier for each referring NP
`ref`	the identifier of the first NP to explicitly refer to the same discourse entity that the current NP refers to. For NPs which mention discourse entities for the first time, `id` = `ref`
`min`	the head of the referring NP (used for scoring)
`type`	the type of the relation between two referring NPs

Table 5.2. Attributes of `coref` element in MUCCS coding scheme

"Markable" is a cover term for those referring expressions that can be marked up according to the guidelines. These expressions are nouns, noun phrases, and pronouns. In the text, markables are enclosed in `coref` tags. The tags delimit the complete phrase including determiners and modifiers, but excluding prepositions. The attributes of the `coref` elements are summarised in Table 5.2.[1] Possessive pronouns are always marked. Named Entities (names, dates, times, currency amounts, and percentages) are also markable, while parts of Named Entities are not. Bare nouns which occur as prenominal modifiers are only marked if they co-specify with the head of another markable or a name or Named Entity. Pronouns referring to propositions or events are not marked, neither are gerunds. Zero pronouns are not marked, either, neither are relations between relative pronouns and the gaps they fill or the NPs they are attached to. A coordination of several NPs is markable, while the coordinated NPs are only markable themselves if they co-specify with another markable. The current version of MUCCS also permits to mark predicating NPs.

SGML II: The MATE Scheme. The MATE project[2] is a pan-European effort to standardise both corpus annotation schemes (Mengel, Dybkjaer, Garrido, Heid, Klein, Pirelli, Poesio, Quazza, Schiffrin and Soria 2000) and corpus annotation tools (Isard, McKelvie, Mengel and Baum Moller 2000). On the basis of the review in (Davies and Poesio 1998), Poesio (2000)[3] proposes a new scheme suited for both dialogue and monologue annotation. It is based on the MUC scheme and strives for conformity to the guidelines of the Text Encoding Initiative (TEI)[4]. The scheme is divided into a core and an extended scheme. The core part only covers relations between expressions which point to the same discourse entity, while the extended scheme allows for a host of other types of relations, most of them inspired by Passonneau's DRAMA guidelines (Passonneau 1996). While the MUC standards collect all information in a single SGML element, the MATE scheme proposes five different elements, summarised in Table 5.3.

`coref:de` is the general element for discourse entities (here: discourse referents), `coref:seg` is used when the referring expression is part of another word, for example, a verb. This covers cliticised pronouns in the Romance languages. For example, the Spanish

[1] In SGML, the structure of a document is described by a set of elements. Elements can include (combinations of) other elements, but elements may never overlap. Information about elements is stored in their attributes.

[2] `http://mate.mip.ou.dk`

[3] (for a summary of that scheme, see Poesio, Bruneseaux and Romary 1999)

[4] `http://etext.virginia.edu/TEI.html`

Element Name	description
`coref:de`	discourse entity
`coref:ue`	items in the visual universe
`coref:link`	type of link between two referring expressions
`coref:anchor`	id of antecedent (embedded in `coref:link`)
`coref:universe`	specify visual discourse universe
`coref:seg`	referring expression that is part of a word

Table 5.3. The elements of the MATE Coreference tag set

word "dígamelo" consists of a lexical morpheme, "diga", the second person singular imperative of "decir" (to say), "me", the first person singular dative pronoun, and "lo", the third person neuter accusative pronoun.

`coref:universe` and `coref:ue` describe situationally accessible discourse entities in dialogue, more specifically items that both participants can (potentially) see. These two elements are based on coding conventions that were developed by Bruneseaux and Romary for task-oriented dialogue.[5] In principle, this approach could be extended in order to specify referents that are part of the hearer's world knowledge (or "larger situation uses", Hawkins 1978)), such as the concepts "birth control" and "drunken driving" or the person "Michael Dukakis" (Massachussetts governor at the time of the WBUR broadcasts analyzed in Chapter 6). However, this is not as straightforward as it seems, since it makes sense to at least distinguish between long term memory and the immediate situation as sources of knowledge about the referent. In principle, something like `coref:universe` tags would be ideal for coding our assumptions about the common ground of communicator and addressee. But their definition is still restricted to task-oriented dialogue. For monologues and everyday conversation, we would need to code other sources of shared entities, such as common memories or common enculturation. How sociologically specific the tags should be depends on the research interests of the annotator.

In the MATE scheme, links between referring expressions are specified in separate `coref:link` elements, which were inspired by the TEI modelling of links. This way, the annotator can not only distinguish between different types of links, but she can also characterise these links more precisely. The antecedents (if there is an identity relation) or sponsors (in cases of bridging) are coded in separate `coref:anchor` elements. The following example is taken from (Poesio 2000, Example 4.15)

(5.1) When do we have <coref:de ID="de_01"> orange juice </coref:de> at Elmira?
We have <coref:de ID="de_02"> orange juice </coref:de> at Elmira at 6 a.m. (Text)

```
<coref:link type"ident" href="coref.xml#id(de_02)">
<coref:anchor href="coref.xml#id(de_01)"/>
</coref:link>
```

For each referring expression, the annotators have to specify how it is linked to the preceding

[5] `http://www.loria.fr/~romary/Documents/index.html`

co-text. Links have two attributes, `type` and `href`. In the extended scheme, which also covers bridging relations, the `type` attribute specifies the type of link. The antecedent is coded in the `coref:anchor` element: The `href` attribute of that element contains the identifier of the antecedent, while the `href` attribute of `coref:link` points to the referring expression itself.

The anchor entity can be difficult to determine when there are several possible anchors which belong to a common script. For example, consider a text where a jail, former prisoners, and a home surveillance system for these prisoners have already been mentioned. Now, the NP "probation officer" appears. Clearly, the NP is not linked to any single of the preceding ones, but to the main discourse topic of the text. Or a text talks about a certain jail, then goes on to mention different subgroups of its prisoners. Are these prisoners inferred from the implicitly evoked set of prisoners of that jail or directly from the frame "prison"?

On the other hand, the attribute becomes indispensable when decisions about bridging are made on the basis of a model of the hearer's world knowledge, because it allows to protocol the basis on which the decision was made. For example, in the pair "the house"—"the door", the door can be identified on the basis of the previously mentioned house on the basis of the connection `hasapart(house,door)` in a knowledge base.

Approaches to Bridging: But the most difficult aspect is surely developing a consistent annotation scheme for inferrables. It is notoriously difficult to develop a consistent annotation scheme for bridging NPs (Poesio and Vieira 1998). Clark (1977) suggests a simple reason: Addressees can build a cognitive bridge between a discourse entity and the addressee's knowledge in so many different ways that no taxonomy will ever cover all of them succinctly. Despite these fundamental problems (discussed further in Section 4.2), non-anaphoric definites are simply too frequent in the real world to be ignored. Annotated corpora both show how often certain resolution strategies apply and help develop new resolution algorithms (Vieira 1998).

In her markup scheme DRAMA (Passonneau 1996), Passonneau identifies several types of bridging references which are summarised in Table 5.4. DRAMA was originally designed for dialog annotation. The set of markables is much less restricted than with MUC. In particular, it also allows to mark VPs as antecedents for referring expressions. In the texts that were used for the evaluation reported in (Passonneau 1997), only a subset of these relations occurred: the possessive/genitive relation, subset, and membership. The results show that precision was better than recall: If a relation is recognised, it tends to be recognised correctly, but quite a few instances are simply overlooked.

Poesio and Vieira (1998) used a much simpler annotation scheme in their study. They defined four classes of definite descriptions based on the classifications of (Hawkins 1978) and (Prince 1981):

anaphoric same head: a definite description with the same head noun occurs earlier in the text

associative: not to be confused with associative anaphora as we have defined them in Section 4.1.1, this category covers all definite descriptions whose heads stand in a semantic relation to their antecedent

larger situation/unfamiliar: this category covers most of Prince's categories inferrable and unused

idiom: the definite description occurs in an idiom

set/subset	*The cookies* were really nice. *Half of them* were filled with cream.
part/whole and physical connection	*The house* is beautiful, and *the garden* is well-kept.
causal inference	*An explosion* shook the neighbourhood. *The noise* was deafening.
propositional inference	*It is so hot.* Well, *this weather* really gives me a headache.
genitive/possessive pronouns	*The boy* carefully ties *his shoes*.
implicit arguments	*The plane* crashed, but *the pilot* survived.
implicit and pseudo partitives	Here is *one of my books*. Where are *the others*?
plurals	*The boy* kisses *the girl*, then *the girl* hits *the boy*, and then *they both* start crying.

Table 5.4. Relations between referring expressions in DRAMA

They later revised the first two categories in order to distinguish between co-referential definite description and cases of bridging. The third category was separated into *larger situation* uses and *unfamiliar* uses. Larger situation definites can be resolved on the basis of what Clark and Marshall (1981) have termed community membership, and unfamiliar definites are brand-new discourse entities that are introduced with enough additional information to make them uniquely idenitifiable, given the co-text.

Poesio and Vieira (1998) found that annotators could only distinguish reliably between first and subsequent mentions. The finer distinctions of the more elaborate annotation schemes could not be annotated reliably. Their annotators also had problems with determining the sponsors of definite descriptions that needed to be processed using bridging inferences.

5.1.3 Evaluation and Conclusions

When we want to study linguistic correlates of entity status in corpora, we run into two problems:

1. We need to investigate the communication process in which our data was produced in order to build the models of communicator and addressee which are central to the management aspects of entity status (c.f. also Chapter 4).

   The problems with labelling bridging come from the fact that the hearer models were not specific enough. But then, developing an adequate knowledge base of the domain you are analyzing is an onerous task even if that domain is relatively small, as was the case for Hahn, Markert and Strube (1996), who have pursued that strategy. Fligelstone (1992) reports that the Lancaster group fended off these problems not by a knowledge base, which would not have been feasible for their corpus (American newspaper text), but by an annotation manual of more than a hundred pages which details solutions for contentious points. Lenat (1995) points out that the knowledge base developed for the CYC-project might be used as a source for world knowledge in anaphora resolution, but so far, I am not aware of any work that uses it.

2. Referring expressions not only specify entities in a discourse model, they also show how speakers perceive and hearers are supposed to perceive these entities (Murphy 1992). They code opinions and beliefs, they evoke social stereotypes, and they delimit social configurations. Theories that are restricted to information processing cannot explain the variation that comes from these aspects. As far as I can see, none of the annotation schemes I have surveyed even begins to address these issues. True, they have nothing to do with the analysis of co-specification sequences, but they are an important aspect of analysing these discourses.

In this thesis, I propose to tackle the two problems outlined in two ways. The first way is to analyze the communication process which formed each text in detail, which gives us the in-depth analysis of limited samples to be found in Conversation Analysis or ethnomethodology (Sacks 1995). This is the path I follow in my analysis of radio news. The in-depth analyses are reported in Appendix A; the quantitative results and a survey of relevant research on the communication situation in Chapter 6. Alternatively, we can completely neglect these variables and measure entity status in a way that makes as little assumptions about the communication process as possible. This is the alternative that I will investigate in more detail in Section 5.4.

5.2 A Source-Based Scheme for Annotating the Givenness of Discourse Entities

This section documents the coding scheme that was used for marking up the radio news texts. As we have seen in Chapter 4, researchers have proposed many competing schemes for describing givenness—or management aspects of entity status, as I prefer to call it here. To annotate them all would be extremely time-consuming. I selected two approaches for further comparison: the cognitively oriented Givenness Hierarchy (Gundel et al. 1993) and a scheme based on Lambrecht (1994) and Passonneau (1996) that codes the source where information about the discourse entity comes from, in particular, the information that we need in order to build the initial description. This scheme is described in Section 5.2.2; some derived taxonomies are summarised in Section 5.2.3. The co-specification sequences themselves were labelled according to a modified version of MUCCS as documented in Section 5.2.1.

5.2.1 Marking Co-Specification Sequences

The basis for marking was the MUCCS scheme, and what was marked were co-specification sequences. Since we have no mechanism for labelling parts of words, we did not label cases where a referring expression has an antecedent in an anaphoric island. This was a particular problem when annotating the German texts, because in these texts, compounds were much more frequent than in the English ones. Reflexives were not incorporated into co-specification sequences. In coordinations, we mark the complete coordination; parts of the coordination are only marked when they are part of a co-specification sequence. Contrary to MUC, we do not distinguish grounding instances. We also do not label NPs that occur in appositions to a head noun or that are arguments of copulas.

Of course, this scheme is far from perfect. In a recent series of articles, Kibble and van Deemter (1999a, 2000) criticise computational linguistic annotation schemes from the point of view of formal semantics. Firstly, they argue that annotations should distinguish between properly co-referring antecedent-anaphor pairs, such as that in Example 5.2 and pairs that merely co-vary, such as that in Example 5.3 where the exact referent depends on the instantiation of the variable that both anaphor and antecedent point to.

(5.2) [The solution]$_S$ we found in our conversation was good. [It]$_S$ works fine.

(5.3) [A solution]$_S$ may emerge from our conversations. [It]$_S$ should work well, given that we are experts.

They also raise the issue of interpreting intensional descriptions whose referent changes during the time that the discourse covers. We briefly addressed this issue in Chapter 4, when we identified the need for an update mechanism. Kibble and van Deemter (1999a) also criticise the notion of co-specification advocated by Webber (1983) and Sidner (1983) because it is not clear to them what specification means, and whether it also includes bound anaphora. On my reading, it does include them.

The real problem is that we have to specify what our annotations are intended to do. If they are meant to elucidate how language can be represented in terms of a formal semantics that relates the discourse to the world, then Kibble and van Deemter have a point. But if the annotations are intended to highlight how repeated pointing to the same entity helps communicators and addressees establish texture, then we need to be more generous. As a criterion for determining co-specification, annotators can use a simple extension of the co-reference criterion: if two referring expressions specify the same entity in the discourse model, be it a variable or an individual, link them. The specification relation links referring expressions with the entities in the discourse model that they access, or, in the case of first mentions, evoke. Evoking is not all or none, as Webber's (1991) research on discourse deixis has shown. Stretches of discourse may only be available as sponsors for a subsequent referring expression for a limited time. For example, discourse deictic pronouns can only refer to regions on the right frontier of a discourse tree; as soon as a region has vanished from that frontier, it becomes unavailable. Eckert and Strube (to appear) exploit this property for constraining the search space in the resolution of discourse deictic pronouns. A similar restriction might be placed on discourse entities that are first evoked by attributive NPs or parts of compounds.

Finally, an interesting feature of the MUC-scheme is that its coverage extends when we change languages. What tends to be expressed by gerund constructions in English, reference to events, and in particular first mentions of events, is expressed in German, thanks to its rich derivational morphology, as a nominalisation.

5.2.2 The Source-Based Scheme

The source-based scheme was developed to track how discourse entities are managed in a given text. The scheme focuses on different initialisation strategies; access routes for entities that are already part of the discourse model were not encoded. This would have required me to commit myself to a particular model of how discourse entities are accessed, and I am still reluctant to do that, as may have become apparent from the discussion in Chapter 4.

The scheme is based on Lambrecht (1994, Chapter 3), with two main differences:

- the category of "inferential accessibility" has been expanded to indicate some frequent types of bridging following Passonneau (1996),
- the category "textually accessible" has been dropped completely.

Discourse entities become textually accessible or *displaced* (Brown 1983b) when they have not been mentioned for a couple of sentences. But when exactly does a subject cross over from the set of "active" into the set of "textually accessible" ones? Should the definition be based on surface form, distance to last mention, or topicality? It appears that "active" and "textually accessible" mark two ends of a continuum; therefore I collapse the two categories and measure the position of a subject on this continuum by distance to last mention.

Another problematic category is *unused*, which encodes assumptions about the hearer's world knowledge. In my analyses of the American English news texts, I assume that the listeners of this radio station, a local station in Boston, Massachussetts, know about the state, the city, and its institutions, but not the people who have positions in these institutions, with the exception of then-governor Michael Dukakis. Furthermore, I assume hearers are familiar with concepts such as "birth control" or "drunken driving". For the German texts, very prominent politicians and well-known companies, such as Daimler-Benz (now DaimlerChrysler) are assumed to be familiar to most hearers. When an institution is referred to by an explicit NP that already appeared as a bare noun modifier, it is assumed to be familiar and inferrable.

The categories in this coding scheme are a superset of the DRAMA categories plus some categories that were introduced in the extended MATE coreference annotation standard (Poesio 2000). They were selected because they are relatively straightforward to operationalise and cover most of the inferrables found in the texts.

Frame inferrability: A new discourse entity d1 is frame inferrable if there is a discourse-old entity d2 to which d1 has a close conceptual connection:

- d1 can be connected to a discourse-old entity d2 using a PP modifier; the resulting NP uniquely identifies a discourse entity
- d1 can be uniquely identified using a relative clause containing d2

The links between d1 and d2 come from the addressee's world knowledge, or, more precisely, from schemata or MOPs (Memory Organisation Packet, Schank 1982) The first criterion is motivated by the observation that if d1 had been introduced as "d1 of/from/at/. . . d2", it would have been brand-new anchored, and not inferrable. The second criterion was added to cover cases like Example 5.4.a, where both "bride" and "church" can be inferred from "wedding". Both connections are best made via the relative clauses given in Example 5.4.b.

(5.4) a) The wedding was really glamorous. The bride wore a diamond tiara and the church was beautifully decorated.
b) *the bride* who got married at the wedding; *the church* where the wedding took place

For this source-based scheme, I defined the intended hearer informally. I specified him in terms of education and political interest ("John Doe", c.f. Chapter 6) and simulated him using my own world knowledge. Two types of frame inferrability which are standard relations

in semantic networks were coded as separate categories: part-whole relations, also known as meronymic or *has-a-part*, which I will call *physical inferrability* here, and set relations, which comprise the well-known *is-a* or hyponym/hyperonym relation, and which I will call *set inferrability*. Following the MATE scheme, I also added *function/value inferrability* for labeling the relation between a numerical value and the variable it is supposed to fill.

Physical Inferrability covers bridging relations based on physical connections and part-whole relations, as in example 5.5. Houses have doors, and doors are fixed to their frames by angles. Physical inferrability can be labelled rather reliably if decisions are based on the physical form of the prototype of the discourse entity which sponsors the new entity. It is also part of both DRAMA and the MATE scheme. Therefore, it was included in the present specification, although it never occurs in the radio news texts because of their restricted domain.

(5.5) a) [The house]$_1$ is in ruins.
b) [The shattered door]$_2$ croaks in the wind.
c) It has not been painted for years.
d) [The angles]$_3$ are covered in rust.

Set inferrability generalises of Prince's (1981) Containing Inferrables. It covers cases where a new discourse entity is an element (Example 5.6.a), a subset (Example 5.6.b) or a superset (Example 5.6.c) of a discourse entity that has already been mentioned in the text (in that example, "bread"). Two types of relations were distinguished, classical `isa`-links (i.e. member-set relations) and subset/superset relations.

(5.6) I bought [lots of bread]$_1$ and some cheese today.
a) I really needed to get some [food]$_2$.
b) [The piece of brown bread]$_3$ was quite nice.
c) But [the buns]$_4$ were barely edible.

Function-value inferrability, taken from (Poesio 2000), is labelled when an expression refers to a value of a function mentioned earlier on in the discourse. The category is extremely rare in the news texts, because there are few measurements and specifications of amounts of money. Example:

(5.7) The wizards pay their cook [2000 gold pieces]$_{GP}$ a month.
They would never have payed [this handsome salary]$_{GP}$ to a bad cook.

Propositional Inferrability: This category handles cases of discourse deixis. It applies when a referring expression refers to a state, event, or process which has to date only been expressed propositionally in the discourse. The antecedents of such discourse deictic expressions are not marked explicitly. Example:

(5.8) Yesterday, [two underground trains crashed in Cologne]$_1$. More than 67 people were hurt in [that crash]$_1$.

Other inferrables: The remaining potential types of bridging, such as causation and plurals, are marked as *other inferrables*. They are very rare in the corpora I annotated.

5.2.3 Coarser Taxonomies

On the basis of the full source-based annotation scheme that I developed in the preceding section, and that is summarised in Table 5.5, I defined four coarser taxonomies with two to four categories:

DISC: discourse-old vs. discourse-new (Prince 1992)
Old entities have already been mentioned explicitly in the discourse, New ones have not.

HEARER: hearer-old vs. hearer-new (Prince 1992)
Old discourse entities are accessible via preceding co-text, world knowledge, or slot in current mental model of discourse, New ones are not.

STAT3: old vs. mediated vs. new (Strube 1998)
Old discourse entities are accessible via the preceding co-text or world knowledge. Mediated ones are accessible via current mental model or via an explicit anchor to the co-text, while new ones are, in Prince's (1981) terms, brand-new unanchored.

STAT4: brand-new vs. unused vs. accessible vs. active
In terms of Figure 4.4, Brand-new entities are unidentifiable, inactive entities are Unused, situationally or inferentially accessible entities are labelled as Accessible, and textually accessible or active items are considered to be Active.

Originally, these subdivisions were introduced in order to test whether, as found by Brown (1983b), prosody only provides rough indications as to entity status, or if there are more subtle correlates.

The dichotomies "hearer old/new" and "discourse old/new" are taken from Prince (1992). Since it is difficult to classify inferrable referents as discourse/hearer-old/new, she introduces inferrables as a third category. I have not followed that move for two reasons: Firstly, the original discourse old/new dichotomy is very easy to derive from co-specification sequences. Secondly, it is interesting to see how far we can get with dichotomies that highlight different aspects of entity status, connection to the co-text (discourse old/new) and connection to the hearer's knowledge (world knowledge, content of short-term store, episodic representation of current discourse). We will assume here that inferrables are discourse-new, but hearer-old.

5.3 Distance Measures

Many researchers talk about entity status in cognitive terms, in terms of newness, accessibility, recoverability, or familiarity. But such descriptions require many inferences about the hearer: What does he know, what can he infer, and what will he forget? In comparison to these rich measures, distance seems to be almost pre-theoretical. Nevertheless, it is widely used for a number of reasons.

First, once we have a discourse with co-specification sequences and segment boundaries, distance measures relative to these segments can be computed automatically. If both sequence and segment annotations conform to reliable annotation schemes, the measures derived from these annotations are reliable, as well.

Source-Based Scheme			Derived Schemes			
Code	Category	Description	STAT4	STAT3	DISC	HEAR
	brand new	unknown to hearer				
BU	*unanchored*	no link to existing discourse entity	BN	new	new	new
BA	*anchored*	link to existing entity	BN	med	new	new
U	**unused**	known to hearer, new to discourse	U	old	new	old
	accessible	initial representation can be constructed on the basis of . . .	AC	med	new	old
SIT	*situation*	. . . the communication situation				
INF	*inference*	. . . link to existing discourse entity X by one of these mechanisms—				
FRAME	frame:	part of script/MOP evoked by X				
PART	part/whole:	physical part of X				
VAL	function/value:	value of X				
ISA	set (isa-Link):	element of X				
SET	set (other):	subset/superset				
EVENT	nominalisation:	nominalisation of VP denoting X				
AC	**active**	already mentioned in discourse	A	old	old	old

Table 5.5. The source-based annotation scheme and derived taxonomies

Second, computing distance information is much faster than labelling any of the categorical taxonomies discussed in the preceding section.

Third, distance-based measures are well suited for typological studies, since they can be defined to be comparatively independent of language-specific categories (Myhill 1992). This allows to compare how distance to last mention affects the form of referring expressions across widely different languages and cultures.

Finally, when annotating large amounts of text, the annotators often cannot construct adequate models of the communication situation because the texts in most corpora are highly de-contextualised. Researchers hardly know who wrote the texts, let alone in which situation and for whom. The Brown corpus is a good case in point. Although we have information about the original authors and publishers, many of the texts mirror the time in which they were written. For those linguists who cannot recollect the Sixties or who are not qualified contemporary historians, the texts can sometimes be difficult to interpret. All that such linguists can do, realistically, is to stick to the surface, to the language itself, annotating co-specification sequences and computing numerical distance measures.

Most distance measures express a "distance to last mention". The general requirements for a distance measure are simple: Distances are defined on a set units into which has been segmented. The distance function maps an arbitrary pair of units onto a natural number n which we will call the distance between these units in the discourse. The function should be a *metric*. This means that it satisfies the following three requirements (Heuser 1993):

1. dist(a,b) $\geq$ 0, with d(a,b) = 0 iff a = b
2. dist(a,b) = dist(b,a)
3. dist(a,b) $\leq$ d(a,c) + d(c,b)

Distance measures can differ in the units they are based on, in the level of granularity, and

in their direction (anaphoric vs. cataphoric). These three aspects will be discussed in depth in sections 5.3.2 (units), 5.3.3 (granularity), and 5.3.4 (direction). But first, let us define that between which we measure distances, mentions.

5.3.1 What is a Mention?

Does a mention need to be an overt noun phrase, or can the NP be omitted or replaced by a verb ending? The answer to this question depends on the language we analyze. For languages with zero pronouns, such as Japanese (Aone and Bennett 1995), Korean (Clancy 1996) or Ancient Chinese (Li 1997), those zeroes clearly count as a mention. In pro-drop languages such as Spanish or Italian, where overt pronouns are marked, verb endings can also count as mentions if the corresponding argument is not coded by a separate NP in the surface form. English and German are neither pro-drop languages nor do they have zero pronouns, if we discount the PRO of binding theory for a moment (Fanselow and Felix 1987). Therefore, we will only count surface referring expressions as mentions:

Definition 5.1 (Mention) *A discourse entity has been mentioned in a sentence iff there is an expression that specifies exactly this entity and if that expression is not bound.*

Givón (1992) also counts omitted subjects and objects in coordinated clauses as mentions. Since in our analysis, such asymmetric coordinations count as one unit, not as two, it does not distort our numbers if we do not count these syntactic gaps as mentions. This decision is consistent with our general strategy: rely on surface form and syntactic analysis as much as possible, avoid semantic analysis as far as possible.[6]

Centering (Grosz et al. 1995) additionally distinguishes between explicit and implicit mentions. Explicit mentions are overt referring expressions, implicit mentions cover cases of bridging. For example, in the second to fourth sentence of Example (5.9), the house itself is not mentioned explicitly, but since the subject NPs can only be interpreted as parts of that house, we can say that "house" is realised implicitly. In fact, it is the backward-looking center of these sentences.

(5.9) The house is really beautiful.
The door is a shiny green.
The roof has been thatched.
The windows are large, with white frames.

Since such bridging references are very difficult to label, and since it is often very difficult to determine the "true" anchor of a bridging reference, we did not count such implicit realisations as mentions.

5.3.2 Potential Units

Distance units can be compared along three dimensions:

[6] Our syntactic analysis relies mainly on mainstream generative grammar, which is another point in which we differ from analysts such as Chafe or Givón.

1. *Is the unit well-motivated linguistically?*
   Does it correspond to the domain of linguistic rules, does it contain complete referring expressions?

2. *Does the resulting distance measure correlate with pronominalisation?*
   If it does, then the measure is useful for answering two important research questions: When can a discourse entity be specified by a pronoun, and which pronouns specify which entities?

3. *Is the distance measure simple to calculate?*
   The base case is simple: If an entity is mentioned twice in the same unit, the distance between these two mentions is 0. If the units do not overlap or nest, the distance between two mentions from different units is just the number of units that occur between two mentions of the same entity. If the units are nested, the calculation becomes complex: we need to represent sequence and inclusion relations between units in a graph, and define our distance measure on such graphs.

Four types of units can be found in the literature: layout units, discourse segments, referring expressions, and clauses.

Layout: In written text, paragraphs and sections are good indicators of discourse structure—if the writer uses them correctly and consistently. Layout structure is also determined by the publication mode (Zinsser 1997) and aesthetic considerations. Many researchers assume that paragraphs are units that can be said to be "about" one topic (e.g. Zadrozny and Jensen 1991). But the reality appears to be more complex. Rodgers (1966) found that topic boundaries not necessarily coincide with paragraph boundaries. Chafe (1994) seconds that argument with sample analyses. Another good example for creative paragraphing is the first section of (Pratchett 1990), reproduced in Appendix A.2. A critical, empirical investigation of the role of paragraphs which compares paragraphs to other models of discourse structure is beyond the scope of this thesis; it requires a specially annotated corpus which I do not have yet.

Referring expressions: This is the smallest unit that still makes sense linguistically. We cannot break down referring expressions into parts of referring expressions that do not specify a discourse entity themselves. Each mention should correspond to exactly one referring expression. Although it is tempting to go down to the word level familiar from many quantitative corpus studies, many referring expressions consist of more than one word. Therefore, we need to stay on the level of full syntactic constituents. You can get around this problem as Biber (1992) did, if you code each referring expression by its head noun. But you lose much interesting information about the internal makeup of a referring expression this way.

A distance in terms of the number of intervening referring expressions tells us how many discourse entities a listener has to construct or access before he needs to access a certain discourse entity again. In order to extend this measure to a measure of something like cognitive processing load, we would need to associate each mention with an activation or construction cost.

Although the unit is linguistically plausible and can be given a cognitive interpretation, it does not predict pronominalisation well. For example, the referential distance between Lucy_l

and she$_l$ in discourse 5.10 is 5, in discourse 5.11, it is 0. Although five intervening referents strain the addressee's memory, it is still possible to refer back to Lucy pronominally.[7]

(5.10) Jim often meets Lucy$_l$ in the little bar with the old-fashioned furniture on the corner of Main Street and Park Avenue. She$_l$ really enjoys talking to him.

(5.11) Jim often meets Lucy$_l$. She$_l$ really enjoys talking to him.

Finally, since referring expressions can be nested within each other, we cannot merely define distance as the number of intervening referring expressions. Let me demonstrate this point by giving a semi-formal definition of distance in terms of referring expressions. First, we need a means for representing nested referring expressions. For that purpose, we use trees with the standard `mother` and `daughter` relations familiar from syntax (Sag and Wasow 1999). `daughter*` is the transitive closure of the daughter relation. We call these trees *RETrees*. The RETrees are ordered according to their position in the discourse.

Definition 5.2 (RETree) *Each referring expression r is represented by a node in a RETree. If a referring expression r_1 is the daughter of another referring expression r_2, then r_1 is the daughter if r_2 in the RETree. Else, it forms the root node of its own RETree.*

Figure 5.1 gives the RETrees for the first two sentences of the text Dayton 2 in Figure 6.5, Appendix 6.4.1.

Definition 5.3 (Distance to Last Mention in Terms of Referring Expressions) *Let r_1, r_2 be two referring expressions, t_1, t_2 the RETrees in which they occur, $distance_T$ a distance measure defined on trees, and $distance_R$ a measure for the distance between RETrees.*

if *(t_1 == t_2)* **then**
  distance(r_1,r_2) := $distance_T$ between the corresponding nodes
else *distance(r_1,r_2) := $distance_T$(r_1, root node of t_1)*
  + $distance_R$(t_1, t_2)
  + $distance_T$(r_2, root node of t_2)

Definition 5.3 is still very general; it is not clear how we should measure the distance between RETrees, or how we should weigh distances within a tree and distances between trees when calculating the overall distance. To see how the measure works, let us consider Figure 5.1 again. Syntactically, it makes sense to assume that the distance between mother and daughter in a RETree is smaller than the distance between two RETrees, since mother and daughter belong to the same complex referring expression. Therefore, we define $distance_T$ as half the number of nodes between a referring expression and the root of its RETree, and $distance_R$ as the number of intervening RETrees. With these definition, the distance between the referring expressions D1

[7] This effect can be explained by two different cognitive mechanisms: salience and frames (Schank 1977). The salience explanation is straightforward: Since Lucy is a human being and in object position, she is more salient than physical objects in adjunct position (Givón 1992, Fraurud 1996). The script interpretation is not less intuitive: All of the NPs and PPs in the example fill a slot in a meeting script (who meets whom where).

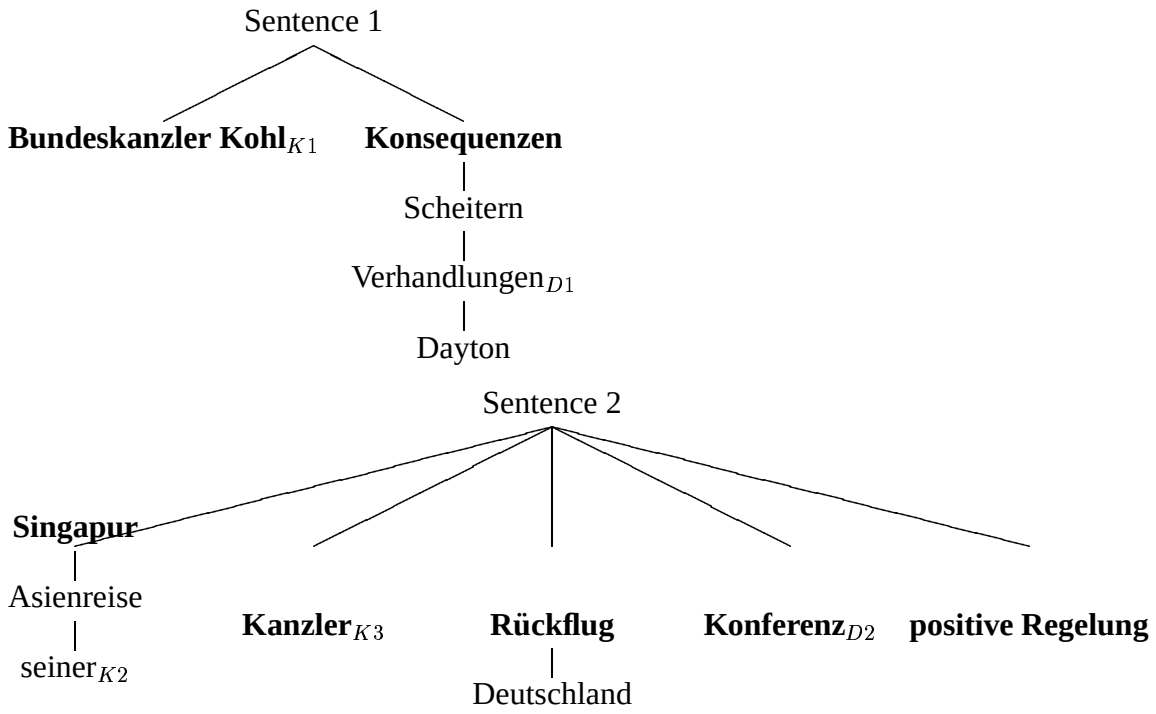


Figure 5.1. Sample analyses of RETrees for the first two sentences in Text Dayton 2. D: Dayton peace talks, K: Chancellor Kohl

and D2, both referring to the Dayton peace talks is 2*0.5 (two nodes between D1 and the head of its RETree) + 4 (intervening RETrees) + 0 (D2 is head of its RETree) = 5. Regarding the references to Chancellor Kohl, the distance between K1 and K2 is 2, and the distance between K2 and K3 is 1.

To sum up, defining a proper distance measure on referring expressions is difficult, and as long as it cannot be connected reasonably well with some notion of processing load, it does not make sense to define such a measure on corpora, except for the sake of experiment.

Discourse Segments: Although purely linear distance measures alone can account for most of the pronouns in a discourse, they sometimes predict a pronoun where a full NP was used. Some cases of pronoun overgeneration can be explained by aspects of discourse structure, such as segment boundaries. Examples and relevant results are discussed in Section 3.3. The problem with discourse segments is that it is difficult to develop reliable annotation schemes which are both language- and genre-independent. As Mann et al. (1992) acknowledge, the segment structure of a text, and in particular the exact rhetorical relations between segments, depend to a large degree on the addressee's interpretation. We also need to keep in mind that there are many other ways of signalling the beginning of a new discourse segment: shifts in time and aspect, discourse markers, explicit changes of location, and so on. Fox (1987) obtained her results on pronouns that referred to persons, and many of her texts can be said to be about persons. No wonder that referring expressions were important cues to structure. It remains to be seen if that result still holds for texts with several main protagonists throughout the complete text, or which are about ideas rather than people.

Discourse segments also present formal problems: As with referring expressions, the distance measure quickly gets convoluted. Again, let me illustrate what I mean by a general, semi-formal definition of distance to last mention in terms of discourse segments.

Let S be the set of segments that a discourse consists of, and let us assume that each word of the discourse belongs to at least one unit. Let $S*$ be a subset of S so that each word of the discourse belongs to exactly one segment $s \in S*$. These units $s \in S*$ are disjoint sets that form a complete partition of the set of words in the discourse. On the basis of these units, we can define a linear precedence relation: a segment $s' \in S*$ precedes another segment $s'' \in S*$ if s' occurs before s'' in the discourse.

The hierarchical organisation of the segments of a discourse can be represented formally by a graph (c.f. Marcu 1997, for such a formalisation of RST). On such a graph, we can now define the distance between two units as follows:

Definition 5.4 (Distance between two Units) *The distance dist(s_i,s_j) between two units s_i, s_j in a graph G is the cost of the cheapest path between s_i and s_j in G. The cost of a path is the sum of the costs of all arcs that are traversed on that path.*

This distance measure specifies a metric on $S*$. The cost function enables us to implement something like forgetting, or the effect of segment boundaries: to traverse a segment boundary is much more costly than to remain within the same segment. The more high-level the segment, the higher the costs. For example, we could define the following cost function on Grosz and Sidner (1986) style discourse trees:

transition within same discourse segment:	cost of 1 per clause
transition from DS1 to DS2:	
a) DS1 satisfaction-precedes DS2	cost of 4
b) DS2 dominates DS1 (nested focus spaces)	cost of 2

Once we have defined how the discourse is to be partitioned into segments, we need to relate the discourse entities to the segments in which they occur, we can define the distance between two mentions as follows:

Definition 5.5 (Distance between Mentions) *Let M(e) be the set of all explicit mentions of a discourse entity e $\in$ E in the discourse. Let Occ be an injective function that associates each mention with the unit in which it occurs. Then, the distance between two mentions m, m' is given by dist(Occ(m),Occ(m'))*

The definition of Occ is still a bit vague: should the unit in which the mention is said to occur the *lowest* such unit in the tree or the *highest*? If we choose the lowest units, we are on the level of $S*$, the smallest segments that are both disjoint and cover the discourse completely. Such segments are clauses, which we will examine in more detail on pages 111f. . If we opt for these units, we can model linear distance effects quite well (Fox 1987, Walker 1998).

If, on the other hand, we promote mentions to segments that correspond to longer spans of text, we need to decide which entities are central in that span, and find a way of promoting mentions of that entities to segments at a higher level of discourse structure. One solution to that problem has been proposed by Veins Theory (Ide and Cristea 2000, c.f. also page 34).

Finally, we need to specify which mentions count as next or last previous mentions.

Definition 5.6 (Next and Last Mention) *Let m ∈ M(e) be a mention of a discourse entity e, and let o = Occ(m) be the unit in which m occurs.*

If there is a m' ∈ M(e) so that Occ(m') precedes Occ(m) and all other m" ∈ M(e) that precede m also precede m', then m' is the **last (previous) mention** *of e.*

If there is a m' ∈ M(e) so that Occ(m) precedes Occ(m') and all other m" ∈ M(e) that are preceded by m are also preceded by m', then m' is the **next mention** *of e.*

I have presented a very general definition of a distance measure based on a precedence relation; in that definition, I have discussed a number of problems that arise when we define a distance measure on discourse structure. Let us now turn to a unit that is much more popular in the literature and makes for much simpler distance measures: the clause.

Clauses: This is the standard unit in the literature. But what is a clause? A language engineer would take a clause to be anything between two full stops (or equivalent punctuation). A semanticist would argue that the adequate unit are propositions, whether expressed by a major clause, a minor clause, or a clause with a non-finite verbal head. But if we are interested in how distance influences the form of referring expressions, the underlying unit should be syntactic, since syntax places considerable constraints on the form of referring expressions (see e.g. Chomsky 1981, Fanselow and Felix 1987).

But what should be the syntactic unit? Along with most researchers in the field, we will use major clauses, because—at least in English and German—many syntactic constraints on the form of referring expressions operate on this level. Some scholars attempt to reduce the phenomena that syntactic binding theory accounts for, such as the use of reflexives, to pragmatic principles or cognitive principles, such as accessibility. For a recent debate, see (Ariel 1994, Levinson 1991, Huang 1993). We have not considered such reductions here, because current binding theory describes constraints on the form of referring expressions that hold within a sentence reasonably well, including gaps. In Example 5.12, one major clause with an overt subject is coordinated with other, subjectless major clauses.

(5.12) And drunken captain Vimes of the Night Watch staggered slowly down the street, folded gently into the gutter outside the Watch House and lay there while, above him, strange letters made of light sizzled in the damp and changed colour . . . (Pratchett 1990, page 7; see also Appendix A.2)

Do we have one or two units here? Since the clauses are coordinated, we assume that the gap in the subject positions of the second and third clause is co-indexed with the subject in the first clause (Büring and Hartmann 1998). Therefore, all three sentences form one unit. This unit is what we call a **Major Clause Unit** (MCU, Strube and Wolters 2000):

Definition 5.7 (Major Clause Unit:) *A* **Major Clause Unit** *consists of a major clause, all coordinated subjectless major clauses where the subject position is co-indexed with the subject of the main major clause, and all minor clauses that are subordinated to any of these major clauses.*

Of course, this definition is far from perfect. For example, paragraph 2.3, Appendix A.2, is replete with sequences of words between full stops that have neither a verb nor a subject,

and uncoordinated subjectless sentences. Some of these "sentences", such as sentence 2 and 3, simulate some kind of repair: Vimes is struggling to find the right words for the city in whose gutters he is reclining. In sentences 5 and 6, the subject is clearly elided. Such cases are rare in the corpus of educated, standard American English that was analysed in (Strube and Wolters 2000), but if our unit definition is to be applicable to all kinds of texts, we will have to mark elided subjects explicitly in the annotations. There are currently no well-validated guidelines for English and German for labelling whether an argument has been elided or not, and for labelling why this ellipsis was possible. Elided arguments are neither part of the MUC specifications (Hirschman and Chinchor 1997), nor of the MATE guidelines (Poesio 2000). Since the empirical work on BROWN-COSPEC focuses on the influence of entity status on pronominalisation, not on the influence of entity status on the form of referring expressions in general, the treatment of ellipsis is left to future work.

Since MCUs form a complete, linear, disjoint partition of a text, we can define a distance measure on them. Let us define an index function which assigns the number i to the ith MCU in the discourse. Then, dist(u_i,u_j) = |index(u_i) - index(u_j)|, which is a metric on **N**, the space of natural numbers (Heuser 1993). The graph that connects the units is a straight line. An unit u_i is only connected to its immediate predecessor and its immediate successor. This simple model has three advantages:

1. It does not assume any specific theory of discourse structure. Instead, it focuses on modelling the strong linear sequence effects that have been observed both in corpus and experimental studies.

2. The costs for each arc can be kept to 1. To determine the cost function for arcs to higher-level units is still an open problem. The results of Fox (1987) suggest that the cost of such an upwards transition should be higher than that of a normal linear transition.

3. MCUs can be viewed as temporal units. On the basis of this reinterpretation, we can now define for each discourse entity a stochastic process that generates its occurrences in a discourse. The distance between two mentions in MCUs is the time that has elapsed between these mentions.

5.3.3 Granularity

When we analyse distances, it is often convenient to reduce the large number of values for distance measures based on small units to a few relevant ones. In most cases, these reductions are theoretically motivated. In anaphora resolution, for example, algorithms frequently operate with the categories "intrasentential", "intersentential: antecedent in previous clause" and "intersentential: antecedent more than 1 clause away". The fewer categories we have, the higher the cell counts in contingency tables, and the more robust the results from statistical tests. The situation changes, however, if we treat distance as an interval-scaled variable—in this case, only such transformations are permitted that preserve the scaling.

A popular distance measure in discourse analysis is the *Referential Distance* (RD) of Givón (1983a, 1992), which has been discussed critically in Section 4.3.6. To recapitulate, the basic unit of RD is the clause. Elided subjects or other arguments of the verb in clauses are also counted as mentions. All distances greater than 20 clauses are mapped to the value 20, which

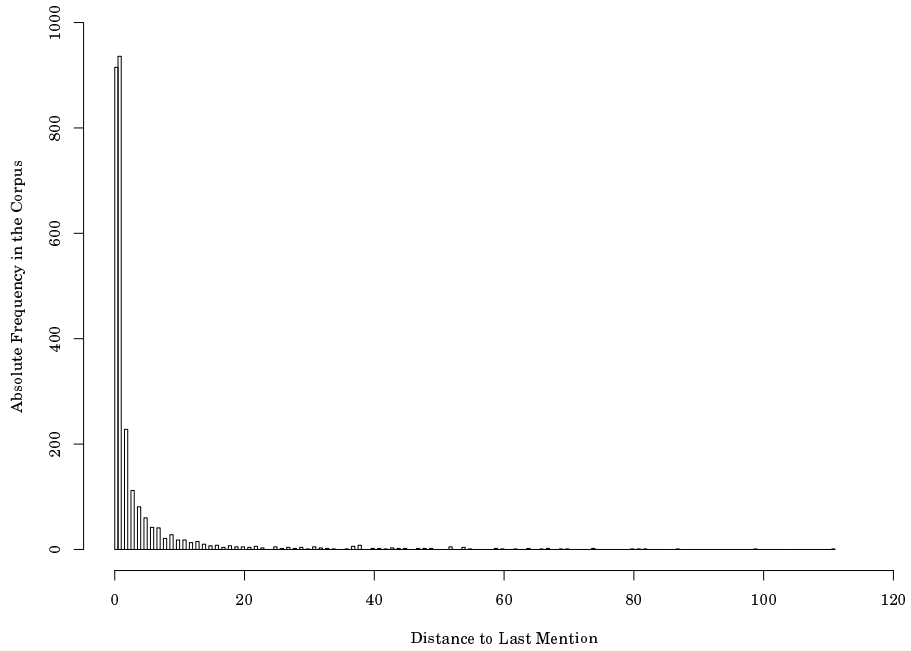


Figure 5.2. Distribution of distances to last mention in the complete corpus

was fixed arbitrarily; first mentions are also assigned this value. I have already discussed some methodological problems with this measure in Section 4.3.5. Here, I focus on the statistical analysis that RD as it stands permits.

First, RD is ordinal, not interval-scaled, because the interval between the cut-off value 20, which is a category, and its predecessor 19, which is still an actual distance, is not well-defined. This is no problem for most analysts, who use non-parametric tests, anyway. A χ^2 analysis is difficult, because in all contingency tables, cells that correspond to distances above 10 will have very few entries. This makes the significance results less valid. Because of the sheer size of the resulting table, it will often be impossible to replace χ^2 by Fisher's exact test. If we treat RD as an ordinal measure, on the other hand, we get access to non-parametric tests such as Kruskal's H-test, which is essentially a non-parametric version of a one-dimensional analysis of variance, and the Wilcoxson and Mann-Whitney tests. Finally, Givón's transformation dramatically skews the distribution of the distance measure. Figure 5.2 shows the distribution of distances to last mention in MCUs for all 12 texts in BROWN-COSPEC without first mentions. We see that the distribution is distinctly reminiscent of an exponential distribution, which could provide the basis of a Poisson process model of co-specification sequences. Figure 5.3 shows what happens to the distribution if all distances ≥ 20 are mapped to 20 (left graph), and when we apply the original definition (right graph). We see that the potential statistical generalisation is lost almost completely. On the other hand, if we do not cut the possible values of a distance measure off at 20, and if we accept that distance measures are just not well-defined for first mentions, we can find a straightforward *parametric* model for distance distributions whose mathematical form can even yield some interesting insights into the linguistic function of entity

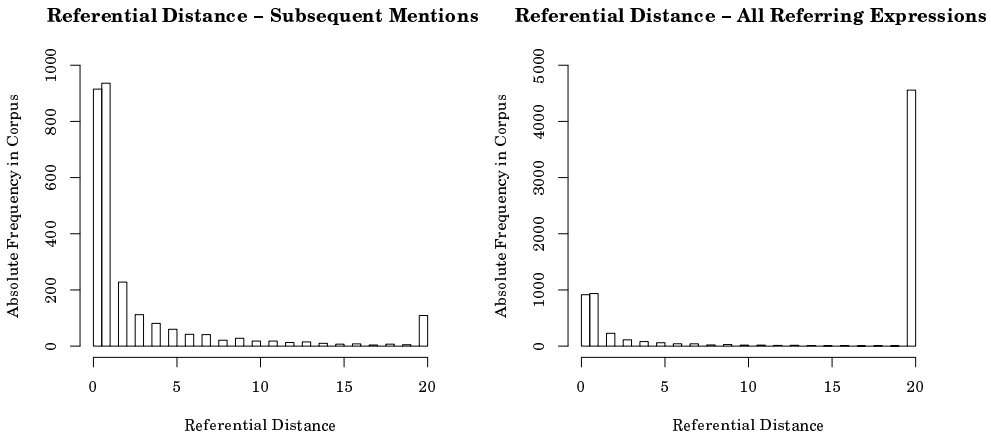


Figure 5.3. Distribution of Referential Distance in the complete corpus. left: without first mentions, right: with first mentions

status. Such a model will be discussed in Section 5.4.

5.3.4 Directionality

Co-specification can be examined from two directions:

- *cataphoric (forward):* For which referring expression is the current phrase antecedent?
- *anaphoric (backward):* What is the antecedent of the current referring expression?

The second question has received far more attention, because it is central to anaphora resolution. The first direction has only been examined sporadically so far, and one of the few researchers to present quantitative results has been, again, Givón. His Topic Persistence (TP) measures how often an entity occurs in the ten units after current one. At first glance, TP appears to be as problematic as RD. Like the first measure, it is based on an arbitrary threshold. Moreover, it is more difficult to compute—at least by hand—than the first one. Taken together with RD, it has a straightforward linguistic interpretation, which is also given by Givón. If TP is high, then the entity is currently topical. If both RD and TP are high, the entity is becoming topical in the current clause. If both RD and TP are low, the entity is losing its current status and is about to become semiactive. With a high RD and a low TP, the entity is clearly not topical. What the measure does not provide are cut-off points which would enable us to state when exactly TP is high enough. In fact, the range and distribution of TP values depend crucially on the structure of the discourse that is being analyzed. For example, TP is lower in short texts with many potential topics than in long texts about a single person. TP may also be lower in argumentative texts, where a single issue is discussed under many potentially relevant aspects. Potentially, the TP values of all the entities in a text tell us much about both the status of the discourse entities that were mentioned and the structure of the text itself.

5.3.5 Summary

We have seen that although distance measures appear to be easy to define and even easier to compute, there are still a number of open research problems. Finding an adequate unit is the first one. We have argued that for written text prose text, MCUs (Major Clause Units) are a good choice, because most syntactic constraints on the form and interpretation of referring expressions operate within major clauses. For speech, the size of this unit is still an open question.

The discussion of Givón's distance measures has shown that distances should either be defined as fully interval-scaled variables, or reduced to a small number of categories with an independent theoretical motivation. Arbitrary cut-off points only obscure generalisations. Therefore, we will use either full distance measures or a four-way distinction between first mentions and subsequent mentions in the same clause, in the previous clause, or earlier. This distinction is motivated by research on anaphora resolution. Since we have only four categories, most contingency tables are densely populated, so that we have a wide range of statistics at our disposal.

5.4 The Stochastic Process of Mentioning

Co-specification sequences document series of mentions. But what is the mechanism that generates these mentions, that generates occurrences of a discourse entity in a text? In this section, I explore what a stochastic model of this mention generating process might look like. Such a model requires large amounts of data, much more than what we have with BROWN-COSPEC. In Section 5.4.1, I introduce the basic statistical model used, the Poisson process, and explore how suitable it is for modelling the data. Then in Section 5.4.2, I propose a more complex approach that the activation level of a discourse entity varies during a text, and how that activation level affects the probability that it gets mentioned.

5.4.1 Foundations

The central data that our model needs to cover are the quantitative patterns that are found in co-specification sequences. In order to model these patterns statistically, we need to translate co-specification sequences into the language of an appropriate probabilistic model. That model is a *stochastic process*: we imagine that each mention of a discourse entity in a text is generated by a random mechanism, and we want to know how this mechanism behaves.

Let us proceed inductively. First we look at the distribution of the distances. For the complete BROWN-COSPEC corpus, that distribution is given in Figure 5.2, page 113. Figure 5.4 gives the distribution of distances for the radio news corpora AUDIX-4 and DLF-RE. We see that the longer a distance to last mention, the less frequently it occurs in the corpus. The frequencies decline exponentially, except for distance 0. This is not surprising: Distance 0 means that the last mention occurred within the current MCU. One reason is that within a MCU, syntactic rules are supposed to influence when and how a discourse entity can be mentioned explicitly. But there is another, more fundamental restriction: propositions tend to be about the relation of discourse entities to *other* discourse entities, not about the relation of discourse entities to

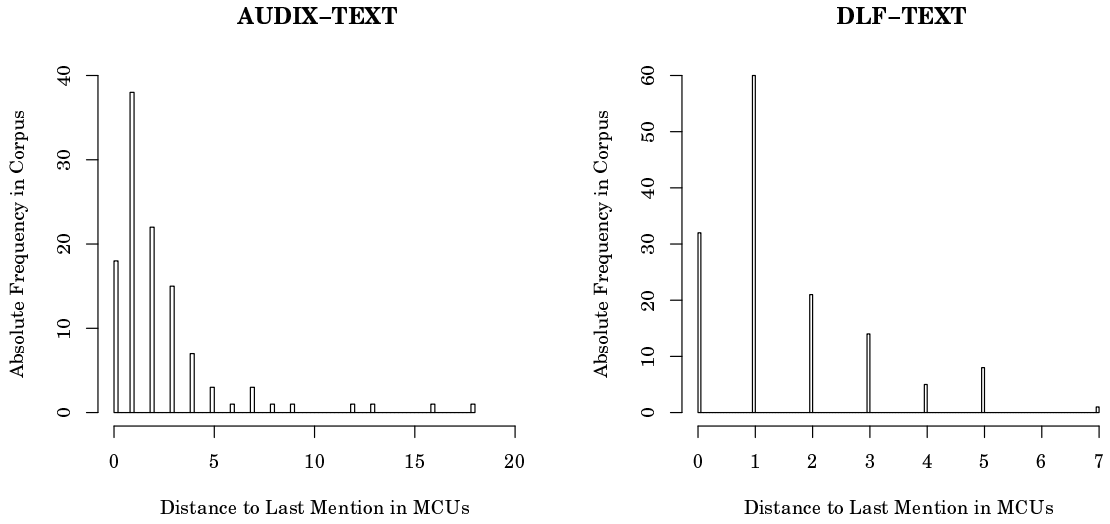


Figure 5.4. Distance to last mention for DLF-RE and AUDIX-4.

themselves.[8] Hence, discourse entities usually only occur once, rarely twice, in the argument positions of a clause.

We cannot just exclude these zero distances from the model, because that would seriously distort the mapping between the statistical model and the domain to be modelled. That domain is the distribution of the mentions of a discourse entity in a text, and we cannot arbitrarily cut it just to suit our model.

The exponential decline of the frequencies from distance 1 onwards reminds one of *Poisson processes* (c.f. Appendix B.5). The events we consider here are mentions of discourse entities, and the random variables X_i describe the distance between two mentions of the same discourse entity. The Poisson process counts the number of times an entity occurs in the text. Distances are reinterpreted as waiting times: the distance to the last mention in MCUs is the "time" the addressee has to wait until a discourse entity is mentioned again. Poisson processes are popular statistical models of language in Quantitative Linguistics (for a recent literature review, c.f. Leopold 1998).[9] But can they be extended to modelling co-specification sequences?

In a Poisson model, we assume that distances between two mentions follow an exponential distribution. Table 5.6 gives the average distances between two mentions for the complete BROWN-COSPEC-corpus and for each genre-specific sub-corpus. For each of these five corpora, I estimated the parameter λ of the corresponding exponential distribution by the inverse of the average distance in the corpus. This is the maximum-likelihood estimator (Lindsay 1995). I then generated a random sample from an exponential distribution with the estimated parameter,

[8] . . . although it is fairly easy to construct a completely narcissistic discourse such as "He first washed himself with his own hands, then held himself tightly in his arms, and decided all by himself to clone himself just to be able to be with himself more often." This MCU contains eight mentions of the same discourse entity, which represents a man who would prefer to remain anonymous for the moment. If we do not count the reflexives, the number of mentions is reduced to three.

[9] Much research in Quantitative Linguistics is limited to word-sized units. In this section, in contrast, I use phrase-sized units, which are better motivated from a linguistic point of view.

Data	all texts		
	all	def.	pro
average distance	3.44	9.79	1.14
KS test statistic	0.2446	0.1453	0.1282

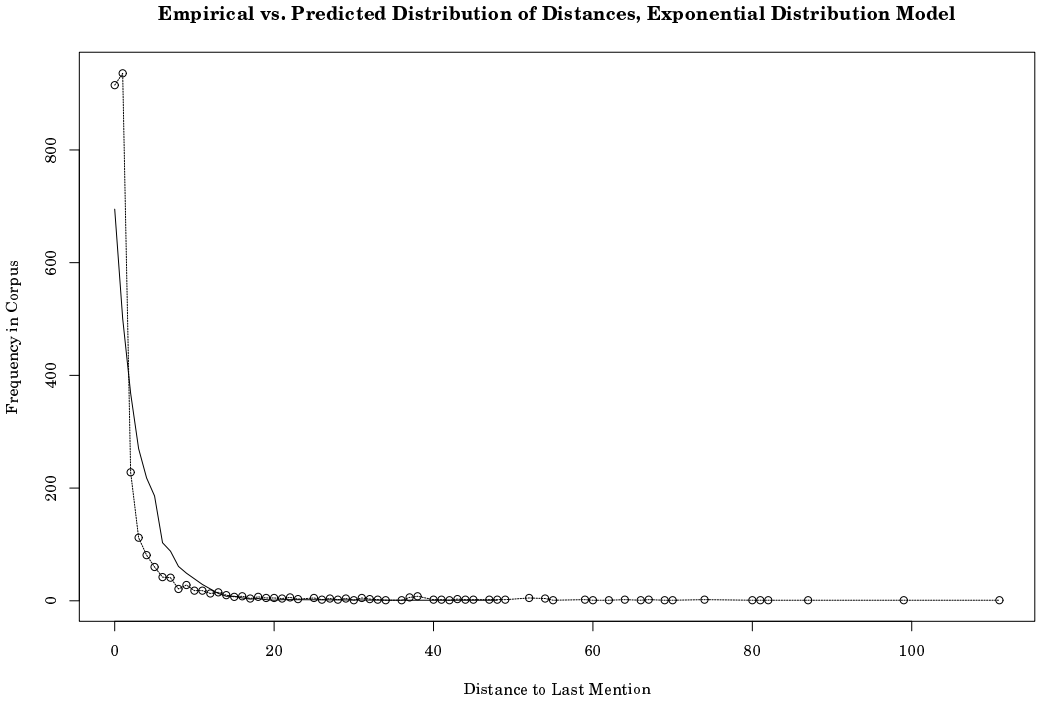
Data	CF			CG		
	all	def.	pro	all	def.	pro
average distance	2.53	4.99	0.84	2.93	4.06	1.29
KS test statistic	0.2285	**0.2571**	**0.0828**	0.1535	**0.1562**	**0.0745**

Data	CK			CL		
	all	def.	pro	all	def.	pro
average distance	2.73	14.88	0.94	4.74	12.33	1.43
KS test statistic	0.2777	**0.1864**	**0.2091**	0.2794	0.1329	0.1447

Table 5.6. Fit of exponential distribution to the data for subsequent mentions. def.: definites only, pro: pronouns only. **bold**: no significant difference between random sample and empirical distribution, criterion: $p < 0.005$, estimates may be unreliable because of a few ties

and compared the sample to the empirical distribution using the non-parametric Kolmogoroff-Smirnoff (KS) test on the BROWN-COSPEC-corpus. The KS test determines for any two samples how likely it is that they were drawn from the same population. The test fails, both on the complete corpus, and on all four genres. Table 5.6 summarises for each corpus the average distance and the value of the test statistic. The lower the statistic, the more likely it is that two samples are from the same distribution. Genre CG, which has the shortest co-specification sequences, comes closest to the estimate. The fit is reasonably good for definites and pronouns, but not for all referring expressions. Moreover, for CL, the distribution never fits well. This indicates that what we need might be a mixture of distributions, not a single one, where the parts of the mixture cover different forms of referring expressions. Moreover, the basic model assumptions might break down for genres where a few discourse entities which dominate large stretches of text.

Figure 5.5 suggests more reasons for these problems. The empirical distribution declines more sharply than the exponential distribution, and has a far longer tail. There is a clear peak around 1, the typical distance for central discourse entities. The problem is not limited to intra-sentential anaphora, where we have already identified syntactic influences that might distort the picture—it concerns the overall shape. Such a pattern cannot have been generated by a stationary Poisson process, or a sum of stationary Poisson processes. The mechanism that generates mentions has to be more complex. The fit is much better when we restrict ourselves to typical anaphoric constructions, such as definites and pronouns. Focusing on pronouns (Figure 5.6), we see that the intra-sentential anaphora distort the fit much more for the complete corpus than for genre CG.



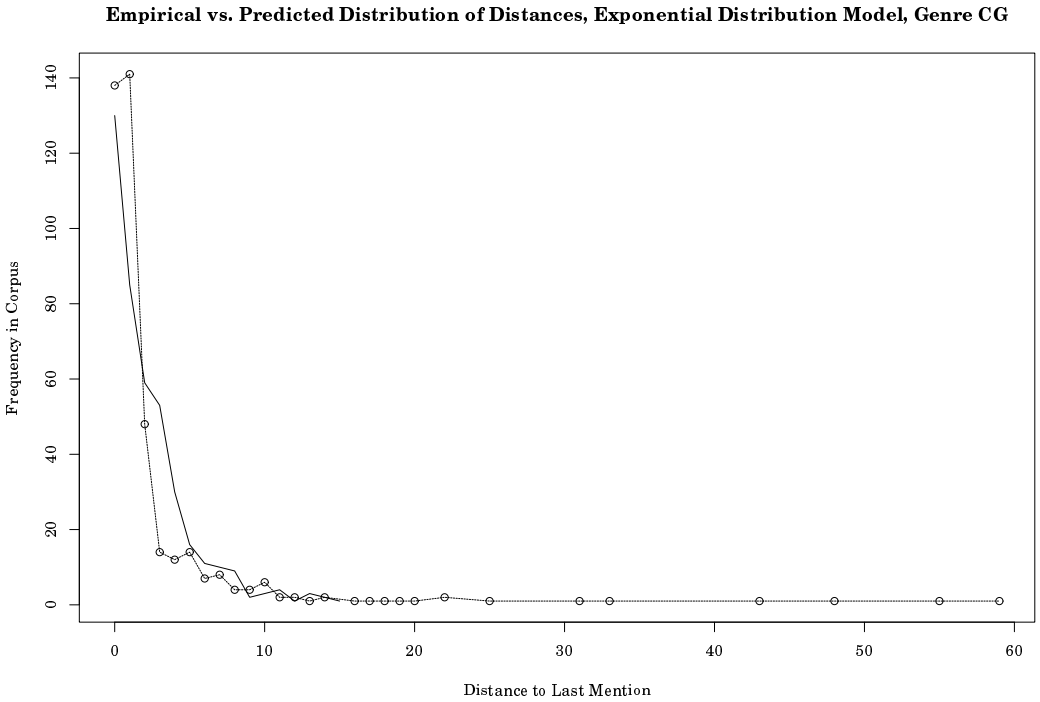


Figure 5.5. Fit of exponential distribution to the data — predicted distances (straight line) versus empirical distances (connected dots). Upper figure: complete corpus, lower figure: Genre CG, best fit

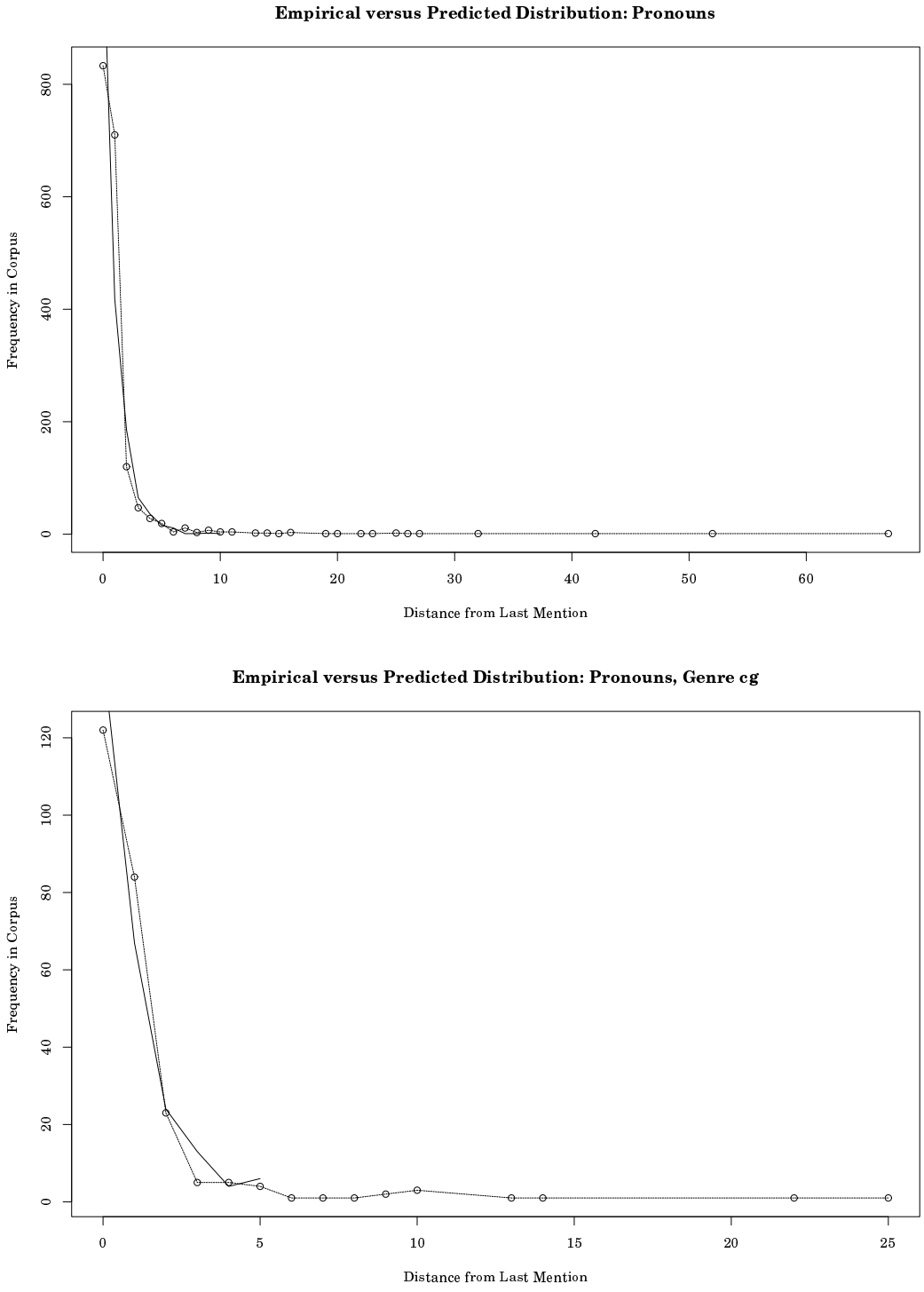


Figure 5.6. Fit of exponential distribution to the data for pronouns — predicted distances (straight line) versus empirical distances (connected dots). Upper figure: complete corpus, lower figure: Genre CG, best fit

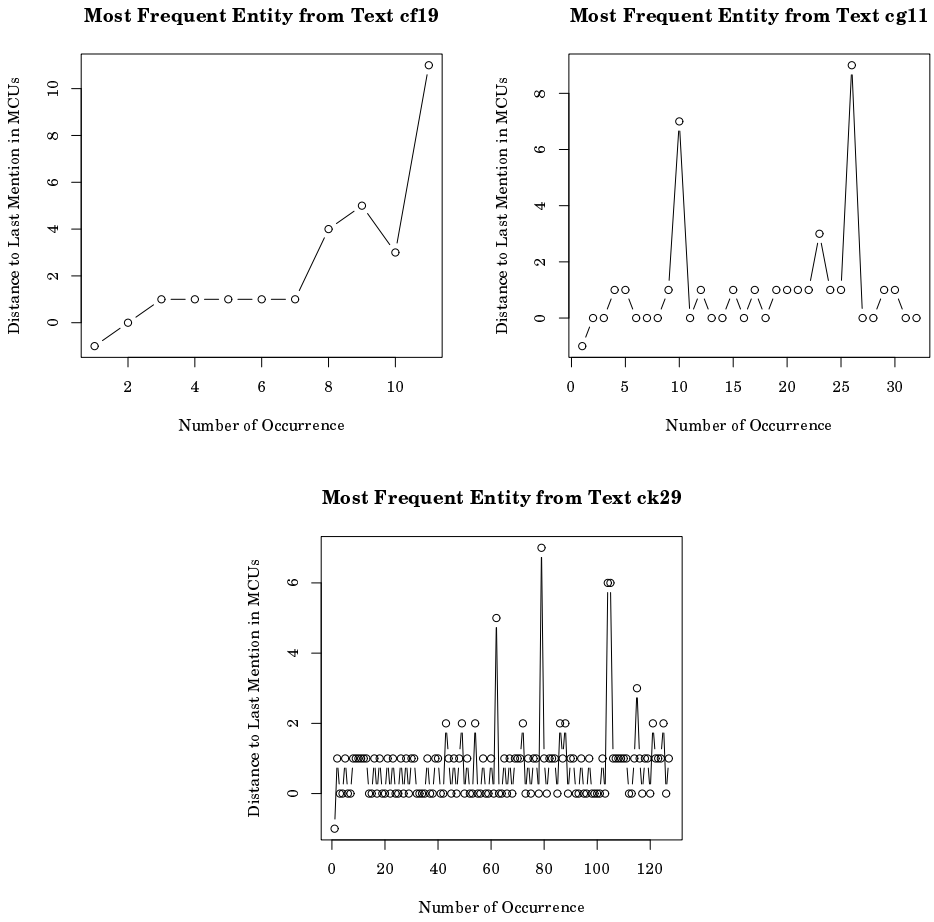


Figure 5.7. Distance to last mention versus position of a mention in a co-specification sequence for three entities. The position is given by "Number of Occurrence". All three entities are the most frequent ones in their respective texts. The first mention has the distance value -1. The graphs show clearly the valleys that correspond to discourse segments where the entities are protagonists.

5.4.2 Active vs. Backgrounded States

What do we make of these results? Why does the exponential distribution fit well for some anaphoric referring expressions, but badly overall? I assume that the real culprit for the bad fit lies in the assumptions that the Poisson processes make, more precisely in the assumption that all distances are independent. Figure 5.7 shows how distance to last mention varies during a co-specification sequence for the most frequent entities of each of the 12 texts. Clearly the process which generates these entities is not stationary. From the graphs, we can distinguish two basic states of those discourse entities:

Case 1: The entity is a central referent. The distance to last mention varies between 0 and 2, which means that the antecedent occurred less than three clauses ago. The entity is

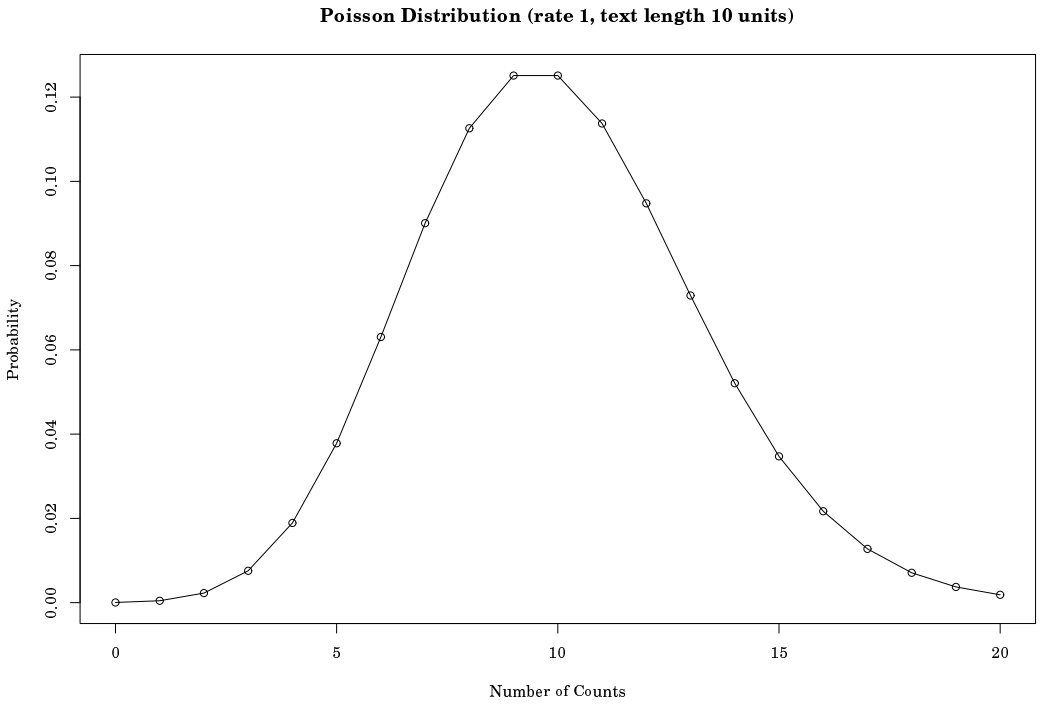


Figure 5.8. Poisson distribution with ρx = 10

highly activated; usually, this high activation persists during several clauses. The entity appears to be some kind of *central referent* in the sense of van Dijk (1980). The rate with which it is mentioned is more or less constant; it oscillates around 1. The process which generates mentions of the entity is stationary. This case can be modelled by a stationary Poisson process with rate 1: In a segment with a length of x MCUs, we would expect the entity to occur exactly x times. If the entity occurs significantly less often than that, it is not central. We can test whether the observed frequency deviates significantly from the expected one using Eq. B.24. Let i be the number of occurrences we found in the corpus, and x the number of MCUs in the corpus, then P(N(x)$\leq i|\rho = 1$) is the probability that we will find i or less occurrences in a corpus of x MCUs, if the entity occurs on average once per MCU. A preliminary evaluation of this heuristic on the BROWN-COSPEC texts indicates that this test is rather conservative. CL06, CK25, and CK29 all have strong central referents, and the number of occurrences in the text is significantly *higher* than what we would expect for a rate of $\rho = 1$ ($p < 0.01$ for the CK-texts, $p<0.05$ for the CL-text). This shows that the "true" rates of the central referents are likely to be higher.

Applying the test only makes sense when the discourse segment is long. Figure 5.8 shows a Poisson distribution with ρx = 10, which corresponds to discourse segments of 10 MCUs in length and a rate of 1 mention per MCU. The cutoff point here is 5: If an entity occurs less than four times in a segment, it is highly unlikely to be a central referent ($p<0.05$). For a length of 5 units, however, this cutoff point drops to 1.

What is the theoretical status of this Poisson model? All the data we have gathered so far seems to suggest that it does not fit well. Even for entities that are central to the whole text, the

model has problems. I compared the distribution of the distances of the three central referents from texts CL06, CK25, and CK29 to the exponential distribution we would expect them to have. None of them fit well—mainly because in the empirical distribution, there are strong constraints on whether an entity can be mentioned again in the same MCU. But as long as we do not have a stochastic model of these constraints, a stationary Poisson process provides at least some reasonable heuristics—the fits for certain types of referring expressions, in particular for pronouns, and for certain genres, in particular CG with its few, brief co-specification sequences are not that bad.

Case 2: The entity is backgrounded. Distance to last mention varies widely. Mentions are spread far apart. If and when an entity is mentioned depends on whether the writer needs it to formulate a particular proposition or make a special point. The process which generates mentions of the entity is non-stationary. The intensity function is now

$$\gamma(u_i) = \gamma(s(u_i), d(u_i), c(u_i), a(u_i)) \tag{5.13}$$

It depends on the state of the discourse model, the locutionary act, and the state of communicator and addressee at time t. To estimate this function from corpora is extremely difficult. One possible workaround would be to estimate the function on the basis of the entities that occur in the current and the previous MCU. If two entities co-occur frequently, then the presence of one entity increases the likelihood that the other will be mentioned. We now have

$$\gamma(u_i) = \gamma(\{e_i | e_j \text{ occurs in the immediate co-text}\}). \tag{5.14}$$

The BROWN-COSPEC corpus is too small and the topics of the texts are too diverse to allow to estimate such probabilities. But corpora of agency stories or news reports about certain topics could exploit these co-occurrences.

Connecting Case 1 and Case 2: The detailed analysis of co-specification sequences in discourse has revealed that the statistical model is not as simple as the alluring curvature of Figure 5.7 suggests. Although stationary Poisson processes give us useful heuristics for entities which happen to play a central role in the segments they occur in, the behaviour of these entities become much more erratic once these segments are finished. We can integrate these two very different cases into a single stochastic model by a simple trick: interpret each case as a separate *state*. For the sake of simplicity, let us assume that these states are linked by a Markov Chain (c.f. Definition B.3, page 274).

In this case, we have two states, let us call them “active” and “backgrounded”. The transition probabilities between those states depend on the discourse structure, more imprecisely on what the communicator plans to say at which point in the discourse. Since the transition probabilities are not stationary, the Markov Chain is not homogeneous. A transition is triggered with each new MCU.

At the next modelling step, we face an important decision: does an entity count as active as soon as it is mentioned, as Chafe would have it (c.f. Section 4.3.6), or should we take a more text linguistic stance and reserve the status “active” to those discourse entities that are central to a discourse segment?

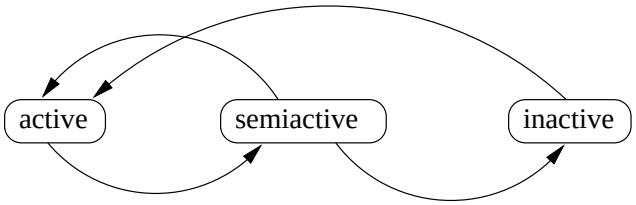


Figure 5.9. Stochastic model of co-specification sequences, Chafe-style

Let us first discuss the approximation à la Chafe. In that case, a discourse entity counts as active as soon as it has been mentioned. If we were to remain faithful to Chafe's ideas, we would also allow a discourse entity to enter the process via the background state. This can occur when an entity is already semiactive because it belongs to what Chafe calls the discourse topic. However, as we have already discussed at length, such co-activation is extremely difficult to estimate from texts. Therefore we will drop it from the present model and assume that all entities enter via the active state.

The probability that an active entity will remain active depends on how often it is mentioned in the following MCUs. The more frequently it is mentioned, the less likely it will plunge back into semiactivity. Discourse entities can also drop out of the process modelled by the chain entirely. This is the state that Chafe would call "inactive". Since inactive discourse entities can be taken up later in the text, we might want to incorporate that state in the model. It will need to be applied when discourse entities are not mentioned for several sections or chapters. A mild case of inactivity is presented by the Pratchett text (Appendix A.2) in the form of the inebriated Captain Vimes — not because he is lying inactively in the gutter, but because the author leaves him lying there all alone for three pages while he reports on strange events at Unseen University. The main difference between the semiactive and the inactive states are the probability of a transition between them and the active state. That probability is much higher for semiactive entities than for inactive ones. Figure 5.9 summarises the model we have developed so far.

While entities can enter only via the active state, they can only leave via the inactive state. For example, in paragraph 3.1, the High Energy Magic Building suddenly enters the discourse, becomes active, but then fades from view quickly as the action centres on the Library. This corresponds to a transition sequence where the entity enters the process in the active state, slips into the background, becomes inactive, and then drops out of the discourse again.

The approach that I have described in the preceding paragraph is in effect nothing else but a stochastic model of co-specification sequences. Such a model has two clear disadvantages:

- The transition probabilities largely depend on the communicator's plan of the discourse, which makes them extremely difficult to estimate from corpus data alone
- The model is difficult to link to the two superordinate states that we found in the data

For this purpose, we need to place tighter constraints on the active state: A discourse entity is active if it can be interpreted as one of the central referents of the current discourse segment, else, it is backgrounded. The chain process for an entity starts when it is first mentioned in the

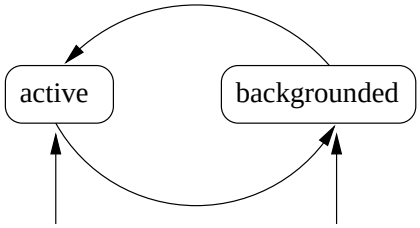


Figure 5.10. Markov Chain model of alternation between active and backgrounded state

discourse. The chain can be entered in both states; the initial probabilities depend on the role of that entity in the discourse. An entity can be introduced into the discourse in either state. For example, the discourse entity corresponding to the Librarian in the Pratchett text (Appendix A.2) is active for quite a few sentences after it has been introduced, while Captain Vimes is not mentioned again after a few sentences (although admittedly paragraph 2.3 is narrated from his point of view).

Both states, the active and the backgrounded one, are not periodic, which means that they can be returned to at arbitrary times. As long as we do not have more sophisticated criteria for determining whether a discourse entity is active, we can resort to a simple test based on the statistical model we posited for Case 1: Let us assume we observe j mentions in the k MCUs that immediately follow the current one. What is the probability of this outcome, if these mentions were all generated by a Poisson process with rate 1? This is the probability that the chain remains in the active state. The larger k, the more stringent our criterion for activity becomes. The resulting transition probability matrix is:

(5.15)

from / to	active	background
active	$p_{aa} = P(N(k) = j \mid \rho = 1)$	$p_{ab} = 1 - p_{aa}$
background	$p_{ba} = p_{aa}$	$p_{bb} = 1 - p_{aa}$

The structure of such a chain is shown in Figure 5.10. To illustrate how the Markov Chain works, Table 5.7 protocols the transition sequences for three discourse entities from the Gemayel text (Appendix A.1), “Jerusalem”, “Gemayel”, and “Arafat”. For the purpose of this sample analysis, our units are paragraphs, and $P(N(k) = j|\rho = 1)$ is calculated on the basis of a moving window of length 9 that is centred at the current paragraph. If an entity occurs only four times or less, it is not active. We see that the discourse entity corresponding to Bashir Gemayel is active throughout the whole text. He is obviously a central referent, but not the topic: that can best be described as the reactions to his death. The case of the entity corresponding to Yasser Arafat shows that the criterion defined above is very coarse. He is clearly the central referent of the paragraphs he occurs in, but since his name is only mentioned three times, which does not quite make our strict threshold of four mentions. The location of Jerusalem, on the other hand, is clearly always in the background, even when, by coincidence, it is mentioned in two paragraphs in a row.

So far we have based our model on what we observed in BROWN-COSPEC, common sense, and analytical categories from text linguistics, in particular, from van Dijk (1980). In order

Paragraph	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Jerusalem																				
active																				
background		X														X	X			
Arafat																				
active																				
background																			X	X
Gemayel																				
active		**X**		**X**	**X**	**X**			**X**	**X**	**X**		**X**	**X**		**X**	**X**			
background																				X

Table 5.7. Transition sequences for discourse entities “Jerusalem”, “Gemayel”, and “Arafat” between active and backgrounded states in the Gemayel text (Appendix A.1)

to estimate transition probabilities from text, we would need a detailed annotation scheme for segmenting discourses and identifying central referents — else we risk circular results. Such annotation schemes are still active fields of research, as we have seen in Sections 5.2 and 3.3. In future work, we plan to annotate the texts in BROWN-COSPEC with discourse structure information; on this basis, the statistical model of the transitions between active and backgrounded state can be further refined.

Evaluation: The results show that the Poisson model, although it is theoretically appealing, makes too strong assumptions about the distribution of mentions. Before we can even think of validating the Markov Chain model of the two states (active vs. backgrounded), we need to solve the problems at the basis. Even if we eliminate all mentions within the same MCU from the picture, we still cannot model the long tail of the distance distribution adequately. In order to cover this massive overdispersion, we will need to resort to more complex point processes (Resnick 1992, p. 332f.). I hypothesise the basic distinction between active and backgrounded states stems from the fact that if a discourse entity is the central referent in a discourse segment, it is very likely to be mentioned at least once per clause in that segment. If this insight is best captured by a model that estimates the time (in clauses) after which an entity will next be mentioned, or if we should switch to a spatial metaphor and count how often an entity is mentioned per segment — that can only be clarified on a corpus which has been annotated with some kind of discourse structure.

Ultimately, we will need to couple the model to both a stochastic grammar, which takes care of the syntactic constraints, and a model of the co-text that takes care of the non-stationarity of the process which generates each mentions. Since the BROWN-COSPEC-corpus is thematically very heterogeneous, it was not possible to explore whether Church’s (2000) non-parametric, collocation-based notion of adaptation might solve the problem that parametricity has given us. As long as we do not have such a superior model, the Poisson-based model gives us a rough, first approximation on which we can build future corpus-based experiments.

5.5 Distance as an Indicator of Entity Status

Distance to next/last mention is defined purely on the text level. Annotating distance does not require a detailed model of the addressee or a deep understanding of the communication situation. This is a disadvantage, because it obviously neglects important aspects of entity status. But it is also an advantage, because addressee models are always subject to much speculation—hence, distance annotation is potentially far more reliable. If an analyst works with data she has not collected herself, from a text type she is not intimately familiar with and where she knows neither typical communicators nor typical addressees, she should assume as little as possible about communicators and addressees.

Basically if you cannot adequately model the communication situation in which a discourse was produced, do not do it. This stance may seem defaitist; and you can certainly get interesting results by throwing all precautions to the winds and placing yourself in the seat of the communication partners, as I have done in Chapter 6. But in the long run this is highly problematic methodologically. If the analyst restricts herself to annotating just those features that she can label reliably, she will very likely not uncover more than the skeleton of the phenomenon she set out to study. But just as the skeleton stabilises the body, reliable findings provide stable starting points for more in-depth analyses. For entity status, this skeleton measure is the position of a referring expression in a co-specification sequence. Explicit mentions evoke and access discourse entities far more effectively than implicit ones. Therefore, it makes sense to start by tracking the explicit mentions. We can get a more balanced picture if we also consider semantic, syntactic, and morphosyntactic properties of the referring expressions in a sequence.

Distance to last mention is no measure of entity status, since that concept is but a cover term that describes how an entity is managed during discourse and the role it plays in texture. It is this role in the texture of a discourse, the *structural* entity status, which we can measure on the basis of co-specification sequences. For example, if a discourse entity is closely related to the topic of a discourse segment, if it is a central referent in that segment, distance to last mention will oscillate between 0 and 2 in that segment. To put it another way, once we have some indication of discourse segment boundaries, we can automatically determine for each segment whether it has a central referent and what that referent is. This procedure also highlights segments where it is not possible to determine a single central referent, as in paragraph 2.1 of the Pratchett text (c.f. Appendix A.2).

On the management side the distance patterns that we have found in the data fall into two states. If the distance oscillates between 0 and 2, the entity is very probably active, otherwise, it is backgrounded. Distances do not allow us to determine when entities are completely deactivated again. This depends on how well they are anchored in the addressee's world knowledge, whether they were primed by another entity that was recently mentioned, on how noticeable the entity was when it was last mentioned, and on how central it was in preceding discourse segments. For example, in the Pratchett text (Appendix A.2) Captain Vimes of the Night Watch is mentioned explicitly only twice, in paragraph 2.2, before he is referred to again in paragraph 3.15 with a pronoun. For readers who have remembered the picturesque gentleman with the alcohol problem and who have wondered when they will meet him again, Vimes will be semi-active. For others he will be inactive and they will have to skip back from page 10 to page 7 in order to find out who this Vimes guy is.

Although the interpretations I have given above appear plausible, they are all based on an

informal analysis of the content of the text. If we do not know how the text begins, how it ends, and who has written it for whom, we cannot really draw inferences about activation or examine the role of a discourse entity in the complete text. This problem is particularly severe with representative corpora such as the Brown corpus or the LIMAS corpus. Many texts are taken from the middle of passages or chapters, others even straddle chapters and sections, so that it is not clear any more whether any of the discourse entities mentioned in part one is meant to surface again in part two of the 2000 word excerpt. We also do not know whether an entity that is mentioned once in the first and once in the last sentence is a protagonist of the segments that precede and follow the extract that has been annotated. Therefore, the results we have presented on the Brown corpus need to be validated on a corpus with full texts or at least the beginnings of full texts. The point I want to make here is that it is indeed possible to develop stochastic models of co-specification sequences, but to collect and annotate a corpus of full texts on which these models could be further refined is definitely beyond the scope of this thesis.

5.6 Summary

In this chapter, we have considered a range of methodologial issues that need to be taken into account in empirical studies of entity status. The core of such studies is the annotation scheme. Good annotation schemes need to satisfy conflicting constraints. They should allow fast annotation of large amounts of data, yet they should be comprehensive enough to cover all phenomena of interest. They should allow high levels of inter-annotator agreement, yet annotators often need to make wild assumptions about how the addressee constructs his interpretation of a text he is confronted with. The annotation scheme should be theoretically well-founded, both from a psychological and from a semantic point of view. No wonder, then, that the ideal annotation scheme has not emerged yet, as Section 5.1 shows.

In this thesis, I explore two ways out of this quandary. For the detailed analysis of a comparatively small amount of data, I developed a detailed scheme that describes how a prototypical addressee might manage the discourse entities that are introduced and taken up again in the course of a brief radio news story. The scheme is described in Section 5.2. Since no detailed hearer model is available to me (this would encode the world knowledge of an educated middle class American or German, respectively), the resulting annotations invariably represent my intuitions about what such a hearer would typically do. The results of this analysis are discussed in detail in Chapter 6.

In Section 5.3, I turn to a radical alternative: distance measures, defined on sequences of co-specifying referring expressions. The requirements for annotation are very simple: Annotators merely have to identify the places where a discourse entity is mentioned explicitly in a text. Defining a suitable distance measure on the resulting sequence of co-specifying expressions is somewhat more difficult. If we have no reliable information about discourse structure, the most straightforward measure is the time-honoured distance to last mention in clauses. In Chapter 7, I explore what distance to last mention can tell us about the mechanisms underlying pronominalisation. In Section 5.4, I investigated whether this measure can help us model co-specification sequences by a stochastic process. This task is more complex than it appears, in particular since we need large corpora for estimating the parameters of such a model and for validating it on test data.

6 Referring in Radio News

This chapter began as a complementary study to research on the given/new distinction in radio news prosody (Wolters 1999, Wolters and Mixdorff 2000). In these studies, we found that entity status is hardly ever signalled intonationally. The first reaction to this result was that these speakers must have been doing something wrong, that I had been investigating suboptimal speech. But there was another option: maybe entity status was already signalled sufficiently well in the text, so that speakers did not need to use additional prosodic cues anymore. When I further investigated the genre of radio news, entity status in radio news turned out to be very difficult to define. The culprit is the rather peculiar communication situation: a web of communicators so heterogeneous that Bell (1991) refuses to reduce it to a single theoretical "speaker"-style unit of analysis, and a heterogeneous audience who merely knows the voice of the person who reads them the news.

These observations lead to two research questions: What is entity status in radio news, and what are its linguistic correlates? Both questions will be addressed in this chapter. Section 6.1 begins with the necessary groundwork from media studies (What is radio news? What does communication in radio news mean?) and explores what entity status in radio news might be. Next, in Section 6.2, I describe the corpora on which the linguistic analyses were conducted, AUDIX-4, WBUR-LABNEWS, DLF-RE and FFH/HR-RE, together with their annotations. In the following two sections, I analyse how entity status is signalled in (radio) news discourse, first quantitatively (Section 6.3), then qualitatively (Section 6.4). I investigate linguistic correlates of entity status in these corpora, focusing on determiner choice, syntactic function, and presence of modifiers. Section 6.5 presents conclusions.

6.1 Communication in Radio News

Many computational linguists are not particularly interested in the genre of text they are working with. For example, in their studies of pronominalisation in pedagogical discourse, Poesio, Henschel, Hitzeman and Kibble (1999) or Henschel, Cheng and Poesio (2000) never refer to the large literature about that type of discourse. Nor do McCoy and Strube (1999) show that they are aware of the lively discussion of media language, even though their corpus consists of newspaper reportages. This is not a serious omission if you are merely interested in describing patterns of language use. But if you start taking the genre you are working with seriously for a change, you discover much that can help you interpret your linguistic observations. Therefore, before I delve into the analysis of the radio news data in Section 6.3, I will survey current and classic results on news language in somewhat more detail than usual in this field of computational linguistics.

This section is structured as follows. In Section 6.1.1, I characterise the genre of radio news, its formats and conventions, then, I survey what gets referred to in radio news in Section 6.1.2. Finally, I sketch picture of the communication process in radio news and draw some consequences for defining entity status in Section 6.1.3. The discussion is necessarily limited: Neither will I discuss in detail the various approaches to (media) communication that have been proposed in the literature, nor do I purport to compare German and American radio. I just want to give you, the reader, a flavour of what we are dealing with here—a detailed study would constitute a thesis in itself.

6.1.1 The Genre of Radio News

According to Swales (1990), a *genre* is characterised by conventions that a community has agreed upon for texts with specific discourse purposes. In the case of radio news, that discourse purpose is: inform the listeners of the radio station about what is going on in the world. The most important convention is that radio news are brief, no longer than three to five minutes. More detailed reports and interviews can be left to other information formats, if the station has them in its program scheme. What is presented as news depends a lot on what the listeners are interested in. Mostly, they get brief overviews of the most important events of the day, information about politics, society, sports, and business as well as service items such as news about petrol price hikes. Stations with teenage listeners tend to spice this mix with the latest from the world of pop music, while local stations often add human interest stories. This means in practice that news stories, even human interest ones, tend to be rather dense, with much information packed into little time.

There are three main formats for news: the "classical" format, with focus on politics, the classical format with sound clips, and the so-called news show, with many service items and few political news (Zehrt 1996, LaRoche 1991b). While both language and presentation of the classical news are rather formal, the news show presents news almost as a form of entertainment; the language becomes more colloquial and the style is relaxed.

The Structure of a News Story: News stories tend to follow the traditional pyramidal scheme "lead, source, background" (LaRoche 1991b, Zehrt 1996, Bell 1991, Lüger 1983). The main news is summarised in the first sentence, while the second sentence provides the source of the news, and the following sentences contain background information. Stories should be structured so that they can be cut from the end, if necessary. This is a common scheme for news reports in general, where a fixed space in time or on paper has to be filled. The content of the story is supposed to answer the question: What happened? or, more precisely: Who did what where, when and how? Writers are supposed to place this information in the first one or two sentences (Burger 1990, Lüger 1983).

If we know how a story is likely to be structured, that story becomes easier to process (Bartlett 1932). van Dijk (1985a) has proposed a detailed prototypical scheme for news, which he applied to the Gemayel text reproduced in Appendix A.1. In his terms, such a scheme is a *superstructure*, a unit of analysis that organises the content-based or intention-based macrostructures, which in turn describe how the propositions a text consists of are organised. The scheme he proposes is given in Figure 6.1. Other analysts have levelled the same criticism against this approach that they have also levelled against van Dijk's theory of discourse structure: Such

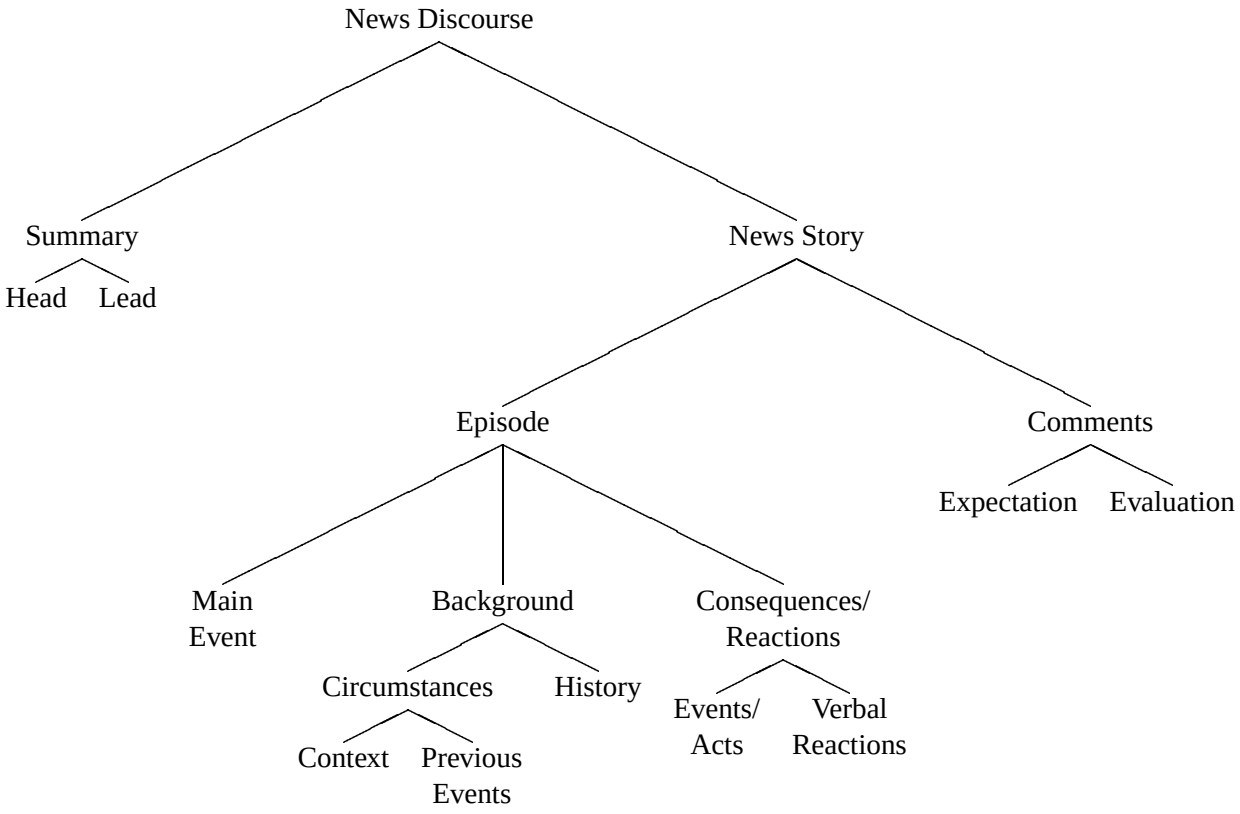


Figure 6.1. Superstructure of news story following (van Dijk 1985a, page 86, Figure 2)

schemes are not flexible enough, and real articles show much more variation than the scheme would predict. Instead of proposing yet another structural scheme, Bucher (1986) developed a set of speech acts that writers use to structure a story. He argues that the foundational level on which addressees construct a text to be coherent is the intentional level. This agrees with our result from Section 3.3: from a communication-theoretic point of view, the basic level of texture has to be intentional.

Bucher's definition of text types ("Textsorten") as *patterns of action* fits quite well with Swales' definition of genre as texts with a common purpose, since actions are *per definitionem* purposeful. He distinguishes between *Meldung* (short news item), *Bericht* (report), and *Reportage* (reportage): short news items inform about facts, reports tell about events and give background, and finally, reportages present the perspective of the reporter. The American English texts are clearly reports, the preferred American form of broadcast news (Stümpert 1991). The German texts, on the other hand, are *Meldungen*, brief news items with as much context as necessary, hence far more difficult to understand and evaluate than reports.

Both perspectives, the intentional and the content-based, complement each other: While van Dijk's superstructures provide a vocabulary for systematically describing the expectations of the addressee of a news story, Bucher's categories are useful for the fine-grained analysis of a story.

The Language of Radio News: Radio news texts are written to be listened to. Listeners cannot spend too much time deciphering the meaning of the current sentence, or they will miss the following one. Neither can they re-read the preceding sentences if they have trouble understanding what is meant. Therefore, the linguistic form should not be too difficult to process—no convoluted sentences with deeply nested NPs or garden path sentences. In reality, there is always a trade-off between the time editors have to write a story, the limited time they have in their programme slot, and the desire to get the story across well. Agency copy (the texts that come in from the agencies) should be rewritten, but due to lack of time, the complex hypotaxes are frequently passed through unfiltered (Haaß 1994) — complete, or rather replete, with nested NPs.

In order to be concise, authors frequently turn verb phrases into complex noun phrases, and insert a semantically rather empty verb such as "take place" into the main verb slot. Most textbook writers polemicise against this practice (LaRoche 1991a, Zehrt 1996, Wachtel 1997, Schneider and Raue 1998) because it makes the resulting text more difficult to understand. But it has its advantages, as Burger (1990) notes. If an event is introduced by a complex NP, referring back to that event is easier than if it had been introduced by a VP. It remains accessible for a longer time, and the initial description probably contains sufficient information for the addressee to connect the discourse-new entity to his previous knowledge and experiences.

6.1.2 What Gets Referred to?

There are two sides to reference in radio news: what is being referred to and how it is referred to. The subject matter of news has been studied intensively by researchers who wanted to find out why some news are reported and others are not. The factors that make an item into a news item influence both form and content of the text. We will deal with that aspect in more detail on page 132 ff. . The linguistic side, the "how", has been studied more globally: researchers have tested the intelligibility of news texts in great detail (a few relevant examples are Früh 1980, Brosius 1990). It is important that the addressee can interpret the news quickly and adequately, not only from a linguistic, but also from a political point of view, since radio is an important source of information for citizens (for controversial discussion, c.f. e.g. Kepplinger 1990, Kepplinger 1999, Noelle-Neumann 1999, Heum 1975, Lazarsfeld and Merton 1948/1971, Lippmann 1922/1971).[1]

[1]Since I leave truth-conditional semantics largely aside here, I will not discuss a third aspect of reference in radio news, the question to what degree the media reality is constructed. Radical constructivists claim that there is no external reality. Instead, everybody constructs her own reality—journalists distill their knowledge into necessarily subjective reports, and their audience uses these reports to construct their own reality as they please (Schmidt 1994, Krippendorff 1993, Krippendorff 1994, to name but a few). On the other hand, realists such as Kepplinger (1993) and Früh (1994) claim that there is indeed a reality, and that journalistic work can be compared to against hard facts. The German discussion in the early Nineties is documented in (Bentele and Rühl 1993); Merten, Schmidt and Weischenberg (1994) provide an introduction to media studies from a constructivist point of view. Although this constructivist approach would fit nicely with recent developments based on Ungeheuer's theory of Communication (Juchem 1998), a detailed discussion of the conflicting points of view would be beyond the scope of this thesis. I am not interested in how the analysed texts relate to the real world; I am interested in how writers use referring expressions to create sequences of expressions that specify the same discourse entity.

What Makes News? Listener expectations are not the only factors that decide which events get selected for reporting and which do not. News agencies flood their subscribers with all kinds of news (Hagen 1995), and many Public Relations managers are busy with creating pseudo-events that are only there to be reported (Kepplinger 1990, Boorstin 1961a). This mass of events and pseudo-events is filtered by the journalists. But according to which criteria?

On one hand, there are the attitudes, values, and ideologies of the editors themselves. A popular metaphor has compared them to *gatekeepers* (White 1950) who decide which news may pass through the gate to publication. On the basis of his interview with "Mr. Gates", White concluded that their decisions are determined crucially by their beliefs, their preferences, and their prejudices. In a classic study, Gans (1980) shows in detail how values and ideologies of the news journalists influence the content of American television news.

News factor theory has strived for more detailed answers to the questions: What do news have in common that are reported by journalists, and what do news have in common that the audience is interested in? In the classic design, news items are classified according to several dimensions, and the researcher tests whether the score of an event on these dimensions will increase the probability that it will get reported by journalists, or paid attention to by the public. The classic list of news factors comes from Galtung and Ruge (1965). They identified twelve factors which can reinforce each other. Since then, researchers have continuously modified and updated that list in a quest for more complete coverage and more reliable categories. Bell (1991) distinguishes between news factors that can be defined on the basis of the events themselves and the actors that take part in them (Table 6.1) and those that reflect the process of writing and editing (Table 6.2). Staab (1990) emphasises reliability of coding: news factors that relate to how close an event is to the audience, be it spatially, politically, economically, or culturally, can be coded much more reliably than news factors that depend on attitudes, evaluations, and choices of persons and institutions. Basically, these difficult news factors describe how people *process* news, given the social group they belong to, their cognitive habits, and their interests. Eilders (1998) investigates cognitive influences on how news are processed and remembered in detail. In both studies, the main coding unit was a news item, and the texts had been carefully selected beforehand from the media coverage at that time.

The study we are dealing with in this chapter is a post-hoc study: it was carried out twelve (WBUR-LABNEWS) to five years (DLF-RE) after the items were originally recorded. In order to get an idea of how the news value of a discourse entity might influence the way that it is referred to, I analysed four texts from DLF-RE, for which I am reasonably familiar with the the social and political context. The results are presented in Section 6.4.1.

A proper analysis in terms of news factors and news value would be far beyond the scope of this thesis. Ideally, I would need to know what addressees remembered of the stories I am interested in when they were first exposed to them, i.e. when they heard them on the radio on the day they were emitted. From this, I could then reconstruct the weight of the news factors. Failing that information, I could plunge into the archives in order to establish the wider context. But such work would merely me allow me to hypothesise about what the addressees might have regarded as news factors—to measure their impact is impossible in a post-hoc study.

Referring Expressions in Radio News: Referring expressions have to be concise, yet precise enough to (re-)activate the background information that is necessary to contextualise the event that is reported. For example, in DLF-RE, Johannes Rau, German *Bundespräsident* in

Events and Actors of News	
Factor	Explanation
CONSONANCE	agrees with audience's expectations
RELEVANCE	affects lives of audience or is close to their experience
PROXIMITY	geographically close to audience
RECENCY	temporally close to audience
FACTICITY	story contains hard facts and figures which are easy to report (who, what, where)
PERSONALISATION	story is about persons rather than concepts
ELITENESS	stories about elite or known (Gans 1980) persons or nations
ATTRIBUTION	quality of the source of a story (elite institution or person)
NEGATIVITY	deviance, damage, death, disaster, conflict
UNAMBIGUITY	clearcut facts, reliable sources
UNEXPECTEDNESS	new, rare, unpredictable (e.g. scientific breakthrough)
SUPERLATIVENESS	the bigger, the better

Table 6.1. News factors according to (Bell 1991, page 155 ff.)—event- and actor-related factors

Production of News	
Factor	Explanation
CONTINUITY	follow-ups to news stories are preferred
CO-OPTION	news is related to a story that draws much attention
PREDICTABILITY	pre-scheduled events such as press conferences
COMPETITION	amount of stories with higher news value during the time span covered
PREFABRICATION	ready-made press releases available
COMPOSITION	editors try to balance different kinds of news items (domestic, international, service, human interest)

Table 6.2. News factors according to (Bell 1991, page 157 ff.)—process-related factors

2000, and *Ministerpräsident* of the German federal state of North Rhine Westphalia (NRW) and deputy chairman of the German Social Democratic Party (SPD) in 1995, is introduced once as "Ministerpräsident Rau" (prime minister, news item 5, 1:30pm, 11/21/1995), once as "Der stellvertretende SPD-Vorsitzende Rau" (the deputy SPD chairman, news item 5, 2:30pm, 11/21/1995). The first item reports that Rau has named a new secretary of state for NRW, the second item reports on Rau's reactions to the leadership crisis in the SPD. This crisis is the subject of another long-running thread that day.

Some news factors are bound to surface explicitly as referring expressions, in particular RECENCY (temporal adjuncts), PROXIMITY (locative prepositional phrases), FACTICITY (mentioning figures, which leads to cases of function-value inferrability), PERSONALISATION (reference to persons instead of institutions), ELITENESS (élite persons tend to be named, others are merely identified by their function). Referring expressions that specify such news factors often convey classical "given" information (especially when a news item scores high on CONTINUITY), but at the same time, this given information is crucial for framing what is new, the news event that is reported. To put it bluntly, cognitive aspects such as accessibility or familiarity may not be as important in the choice of referring expressions as news factors.

Looking at sample news text 6.6 on p. 169, we can immediately identify several referring expressions that specify news factors: We have a report about an elite person, the news item is personalised (both times: Rau), and it continues an ongoing drama (internal social democratic strife), which is referred back to by a nominalised verb. It does not really make sense to ask whether any of the discourse entities that these expressions evoke are given or new. Listeners may have forgotten about the social democrats' troubles, or they may not immediately remember Johannes Rau, although he already was a prominent figure in German politics at that time. (At the time of writing, he is President of the Federal Republic of Germany.) But what counts in this context is that these are the reasons why Rau's reassuring statements are important enough to become news.

So far, most research on the linguistic realisation of these news factors has focused on the question: How are the "news actors", the protagonists of a story, labelled? (c.f. e.g. Kniffka 1980, Bell 1991, Jucker 1992, Jucker 1996). These labels depend a lot on the journalists' attitude towards the actors, as well as on editorial policy and political correctness.

When describing the style of a particular news medium, the form of referring expressions is an important variable. Jucker (1996) shows that up-market, down-market and moderate newspapers differ in the way they introduce news actors. While up-market papers use the proper name plus an indication of the social role as a NP modifier (e.g. "Mr. John Major, the Prime Minister"), down-market papers drop the article altogether and begin with the function or a suitable epithet ("redhead Fergie", "Prime Minister Major"). Subsequent mentions may again be bare NPs, peppered with suitable epithets that convey additional information about the news actor. Burger (1990) notes that verbatim repetitions of a definite NP tended to be avoided in subsequent mentions. This custom has raised the ire of influential textbook writers, such as LaRoche (1991a). They consider that using synonyms or hypernyms or smuggling new information into an anaphoric definite NP where a repetition or a pronoun would do only serves to confuse the unsuspecting audience. Schneider and Raue (1998) put it this way:

> Wechsel im Ausdruck ist bei Verben, Adjektiven, Präpositionen vorzüglich, bei Substantiven meist unmöglich und absolut nicht erstrebenswert. Darin ist sich die Verständlichkeitsforschung mit den meisten Stillehrern einig. Die meisten Journalisten sehen das anders.[2]
>
> (Schneider and Raue 1998, page 194)

For example, take the case of a couple of young English soccer fans who vent their frustrations a bit too loudly in Belgium. When this event is reported in the news, the referring noun phrase "some soccer tourists from the United Kingdom" conjures up other associations than the phrase "English hooligans". Both can be a nuisance to Belgian police forces, but the second noun phrase evokes related experiences with fanatic Brits more quickly than the first. Scripts and stereotypes are called up which influence the story into which the rest of the news about the soccer incident will be embedded. Note also the difference in determiner choice. Both NPs are referential in the sense of (Gundel et al. 1993); they identify a group that we will hear more about later in the news item. The form of the second NP, however, is faintly reminiscent of the bare plural that English uses for generics.

6.1.3 Entity Status in Radio News Communication

All linguistic theories of givenness operate with a communication model which resembles the classical information-theoretic model: A speaker communicates verbally with a hearer, language being the code both share, over a channel constituted by voice and ear. Within this model, most researchers have focused on the intricacies of the speaker's and the hearer's cognition, and on the conditions for using the linguistic code feliticiously. When we now want to examine entity status in radio news, we need to translate this model onto radio news. It would lead us too far afield to summarise even part of the relevant research in communication theory and media effect research; for reviews, see e.g. (Bell 1991, Noelle-Neumann 1999, Merten 1994, Schulz 1999, Schenk 1999). Here, I merely highlight three approaches to show how what a difference the perspective can make.

The classic picture of the media communication process was painted by Lasswell (1948). His famous five categories are given in Figure 6.2. They are mainly intended to point to fields of analysis. What I am doing in this chapter, the analysis of referring expressions in radio news, would be "content" analysis from Lasswell's point of view. When this analysis is spiced up with a few conjectures on the effect of some linguistic choices on the intelligibility of a news item, we move into the field of effect analysis, and when we speculate why an editor might have preferred a certain referring expression over another, we formulate hypotheses about control analysis. Berger (1995, Chapter 1) and Pürer (1998) review some other, more formal, models of (media) communication that are also heavily indebted to information theory.

Krippendorff (1994) presents a fundamentally different perspective on media communication. He distinguishes three perspectives:

- the *theory of communicative competence*, which explains how individual audience members maintain their cognitive autonomy, how they select the news and how they integrate

[2] To change expressions is desirable with verbs, adjectives, and prepositions, but often impossible with nouns and absolutely undesirable. That is a point intelligibility research and most style teachers agree on. Most journalists disagree.

who	says **what**	in **which channel**	to **whom**	with **what effect**
control analysis	content analysis	media analysis	audience analysis	effect analysis

Figure 6.2. Lasswell's (1948) view of the mass media communication process

the information into their construction of reality. Research on the effects of news media has collected solid evidence for the cognitive autonomy of the audience, starting with the classic result of Lazarsfeld, Berelson and Gaudet (1944). They examined how U.S. citizens used the media in order to get informed about the candidates in the upcoming presidential elections. They found that the increased attention they paid to media reports only strengthened the voters in their previous convictions—no matter whether Democrat or Republican. Ruhrmann (1989, 1994) has proposed a selection-based analysis of media effects. In a large-scale study, he found that the audience members actively construct the context in which they embed the news they hear.

In Ungeheuer's terms, the theory of communicative competence describes how individuals integrate the information from the news into their personal experience theory (PET). If we were to define entity status in terms of Krippendorff's constructivist approach, we would need to embed it here.

- the *theory of pattern transmission*, which explains how individuals coordinate their actions verbally, and how certain patterns of actions are transmitted from one place to another. This theory would provide us with the basis for describing how different aspects of entity status are coded linguistically.

- the *theory of communicative authority*, which explains when individuals sacrifice their cognitive autonomy in order to admit outside influences. The critical analysis of how news actors are labelled would fall in the realm of this theory.

Früh (1992b) proposes a model which potentially integrates all aspects of mass media communication: the dynamic-transactional model. The model is *dynamic* because communicator and addressee (in Früh's terminology: recipient) change during the communication process, and *transactional* because both sides interact. Früh and Schönbach (1982) emphasise three aspects of the model: It does not artificially separate dependent from independent variables; the ability to process news interacts with the interest in news; and temporal changes in the effects of news coverage, both quantitative and qualitative ones, can be modelled. Schönbach and Früh (1984) distinguish two types of transactions: Inter-transactions between communicator and addressee (in their terminology: recipient), and intra-transactions, which occur within the addressee or the communicator and which are mainly cognitive. Figure 6.3 shows the basic structure of the model (which becomes far more complex in actual studies). The dynamic model predicts that (news) texts do not have "a" meaning, which is constant across persons or even for the same person. Früh (1992b) found that when readers process a text, they soon activate schemata which guide the way that they interpret the propositions that follow. Readers attempt to fit incoming information into the framework they have selected, and are extremely reluctant to revise their choice of framework when they come across information that is not consistent with it.

In terms of the dynamic transactional model, entity status protocols the intra-transactions in the addressee which take place while he interprets the incoming referring expression. The task

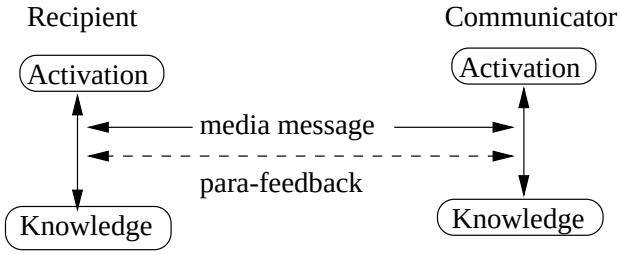


Figure 6.3. The Dynamic-Transactional Model of Media Effects. (Früh 1992b, Figure 3, page 53)

of the communicator is to make this interpretation possible, using strategies that both communicator and addressee know. This is the main inter-transaction here.

The results collected in (Früh 1992b, Früh 1994) have interesting consequences for the study of entity status. When an initial description of a discourse-new hearer-old entity is constructed, a complex web of memories (or, less cognitively, the past experiences and connections to other signs) is activated. This web contains expectations which can be as detailed as scripts. These expectations guide how other discourse entities will be constructed, and what information about them will have been integrated into the model when the memory of the original phonological encoding has faded away. These expectations control what the audience pays attention to, and how the activation of the other entities in the current discourse model increases or decays. Finally, the more detailed the expectations and experiences of an addressee are, the easier it will be for him to interpret a given news item as coherent, even though it lacks familiar surface cohesion markers such as recurrence or pronominal anaphora. For example, text Dayton1 (Figure 6.4) consists almost completely of all-new sentences. It is nevertheless highly coherent: Sentence 1 evokes a script that sentences 2 and 3 elaborate on. This script allows bridging references such as “eine Einigung” (sentence 2) and synonyms such as “Delegationen” (*delegations*) / “Vertreter der Konfliktparteien” (*representatives of the sides of the conflict*). Which discourse entities are familiar, which ones are easily accessible, and which ones are activated now depends to a large extent on the addressee, the scripts that he knows and the encyclopedic knowledge that he has. Without a good addressee model, we cannot annotate entity status. Well, no problem here, you might contend. We just assume—whom? Which of the two types of audience that Früh (1994) found will we assume? The politically interested, educated person who remembers stories well and follows current news out of interest, with well-honed opinions and a large background knowledge? Or the normal guy, let’s call him John Doe, who might follow a story line when it catches his interest, but forgets it as soon as it drops out of the news? I chose John Doe: He knows his country, the main political players, the main stories of the day, but not much more. I chose him because he is easiest to reconstruct for non-contemporary historians after five years have elapsed between the recording of the news items and their analysis.

6.2 The Corpora

Two main corpora were used in both the statistical analyses and the machine learning experiments: the Stuttgart Radio News corpus (SRN) and the Boston WBUR radio news corpus. Both corpora contain read speech from several speakers and have been used extensively for both prosody research and building prosodic modules for speech synthesis systems. The speech in these corpora is based on the educated standard varieties of German (SRN) and American English (WBUR). To examine in phonetic detail the degree to which the varieties in these corpora do indeed correspond to the language norm is beyond the scope of this thesis, especially since this would have required a more detailed discussion of language norms. None of these corpora were collected by myself. What I added to them were linguistic annotations of the news texts themselves, discussed in section 6.2.3.

The corpora are small; they are just the size a single annotator can handle on her own if the annotations are somewhat complex. They are also not a representative sample of German radio news texts, let alone of American ones. However, since both SRN and WBUR are frequently used in the speech technology community, I judged that knowing a little more about the texts in these corpora might not only benefit my research on these corpora, but that of others as well.

The analysis is based mainly on two radio news corpora, WBUR-LABNEWS and DLF-RE, with a combined length of 4093 words, containing a total of 1334 referring expressions. Since the texts in both corpora come from only one source each, we cannot be entirely sure that the peculiarities we find are not merely due to in-house style. Therefore, results on these corpora are supplemented by analyses of two further corpora, AUDIX-4, and FFH/HR-RE. These two corpora roughly match the first pair in size: They contain a total of 4047 words and 1034 referring expressions. While WBUR-LABNEWS consists of radio news reports from a Boston National Public Radio Station, the texts in AUDIX-4 are unedited agency copy. In DLF-RE, we find transcriptions of news flash items from the prestige German radio station Deutschlandfunk, while the texts from FFH/HR-RE are taken from the original manuscripts for news flashes for FFH, a commercial regional radio station, and HR, a public regional station. The annotation of WBUR-LABNEWS and DLF-RE is much more detailed than that of AUDIX-4 and FFH/HR-RE, because these corpora were also used for the prosody studies reported in (Wolters 1999, Wolters and Mixdorff 2000). Together with these studies, they present a detailed picture of how different linguistic means — choice of referring expression and accentuation— are coordinated to signal the current status of a discourse entity. AUDIX-4 and FFH/HR-RE merely serve to validate the main results of the linguistic analysis. Therefore, they were only annotated with referring expressions, co-specification sequences, and sentence boundaries. Each of the referring expressions in these corpora was also annotated with a rough description of its form, which is given in Table 6.7, which is reproduced here for convenience. Finer distinctions, such as between indefinite articles and cardinals, or between the definite article and possessives, are not made, because there is not enough data in the corpus I annotated. Note that in Table 6.7, the definitions of definiteness and indefiniteness are no longer purely morpho-syntactical (presence of the appropriate article), but approach the abstract categories of definiteness and indefiniteness as proposed by Lyons (1999).

6.2.1 American English: WBUR and AUDIX

Code	Topic
J	new judge to be named on Massachusetts Supreme Court
P	prisoners serve their sentence at home thanks to an electronic surveillance system
R	stricter laws against drunken driving
T	school-based health clinics against teenage pregnancies

Table 6.3. Overview of texts in WBUR-LABNEWS

The Boston Radio News Corpus: The Boston Radio News Corpus (Ostendorf et al. 1995) consists of speech from seven radio newscasters from WBUR, a Boston public radio station, licensed by the Corporation for Public Broadcasting.[3] Although the license for operating WBUR is held by Boston University, WBUR is a professional station: In 1996, it had 103 full- and part time employees and 40 student employees. It has been primarily a news station since the early Eighties. A large part of WBUR's programming is supplied by National Public Radio, the oldest and largest public radio network in the US (Bliss 1992). This contract started in 1980. Today, it also receives material from the BBC and Public Radio International.

WBUR's core audience lives in eastern Massachusetts and southern New Hampshire. Its listeners tend to belong to the upper or the upper-middle class. They tend to be affluent, educated white-collar workers. According to the station's own publicity, WBUR is a primary source of information about New England on public radio. From this and from the number of prizes the station has won over the years, I conclude that it is a high-quality well-respected station. I conjecture that the texts in WBUR-LABNEWS were originally written for *Morning Edition*, a popular NPR news show with time slices for local correspondents.[4]

The corpus consists of recordings from seven speakers. Most of the material was recorded during broadcast except for four stories, which were recorded in a studio. These stories constitute the WBUR-LABNEWS-corpus. The newscasters read them twice, first in non-radio style, secondly in radio style. The corpus has been transcribed by hand from the original recordings. Table 6.3 presents an overview of the texts and their contents.

The Audix Corpus: The AUDIX corpus (Hirschberg 1993, page 4) also represents news speech, albeit of a different kind than the WBUR data. It consists of ten Associated Press (AP) news stories which were read by a female professional newscaster under laboratory conditions in radio news reading style. There were no disfluencies, because disfluent productions were re-recorded immediately. Four of these stories were selected for analysis. Table 6.4 gives an overview of these stories and their length.

As we will see below, there should be a clear difference between the language of these two corpora. Agency reports are usually not designed for being read aloud, while texts for newscasts should be. A simple comparison of sentence length between the two corpora shows that while the average length of a sentence in the four labnews texts is 19 words, the average length of a sentence from AUDIX is 23 words (lengths rounded to full words; data from Hirschberg 1993, page 5).

Radio news editors not only often simplify the syntax and the vocabulary of the agency

[3] All information on WBUR, if not stated otherwise, are from the WWW pages `http://www.wbur.org`.
[4] Internet home page: `http://npr.org/programs/morning/`

Code	Topic
2	bribery case against Teledyne electronics
4	broadcasters argue against Cable Act
5	in-flight problems with a Concorde
6	new products from Sun Microsystems

Table 6.4. Overview of texts in AUDIX-4

texts. They also tend to cut agency texts, sometimes quite dramatically, as the examples in (Bell 1991, Chapter 3) demonstrate, because they have to filter the information they receive from the agency to fit both their audience and the programme's time schedule. This observation is underscored by the fact that the average length of a text is 5 minutes (Hirschberg 1993, page 5). It is highly unlikely that editors would subject their audience to 5-minute reports written in the original agency style.

Nevertheless, agency material is the basis of both the content of most news bulletins and the source of most of the sentences that are broadcast. Thus, we would expect that the task was not too difficult for the speaker, and Hirschberg (1993, page 4) observes that the speaker produced normal news reading style. In sum, the AUDIX corpus does not really present a faithful picture of radio news style. Rather, it reflects a typical kind of task for reading machines: reading a complex text out loud so that it is easy for the listener to understand.

6.2.2 German: DLF, FFH, HR

The Stuttgart Radio News Corpus The Stuttgart Radio News Corpus (SRN; Rapp 1998) consists of radio news bulletins aired by the German radio station Deutschlandfunk in 1995 on the full and on the half hour. Deutschlandfunk (DLF), now a part of DeutschlandRadio, is a state-owned non-commercial radio station. It can be received in all parts of Germany and is one of the most prestigious stations in the country.

The DLF news were read by three different professional news readers, two female speakers and a male speaker. The data were recorded directly from the original digital broadcast on two days, July 28 and November 21. The researchers did not have access to the speakers' manuscripts. Bulletins on the hour tend to be longer than those on the half hour. This is almost entirely due to the long bulletins around lunch time (12:00, 13:00, 14:00), a prime listening time, and at 16:00. Table 6.5 gives an overview of the texts and their topics. Only comparatively long news items were selected for analysis. Some texts share the same topic, but were updated according to new developments or rewritten. All texts were read by the same trained speaker, whose prosody has been analysed in e.g. (Wolters and Mixdorff 2000, Mixdorff and Fujisaki 2000, Mixdorff 2000).

The FFH/HR Radio News Corpus This corpus is published in (Haaß 1994, Appendix page 108 ff.). The Hessischer Rundfunk (HR) is the public radio station of the federal land of Hesse, while Radio FFH is the first commercial radio station of that land. FFH presents a news bulletin 5 minutes before the full hour, while HR has a news bulletin at the traditional times. Haaß (1994) compares the two news styles. The corpus will be used mainly to check hypotheses

Day	Hour	News Item	Position / Topic
July 28	12:00	1	attack on Bosnian town of Bihac
		4	Mazowiecki steps down as UN special emissary to Bosnia
		6	French atomic bomb tests in Pacific
		8	Castro wants to stay in power in Cuba
		9	Kurdish protests and hunger strikes
		10	Social Democrats' comments on tax compromise with Christian Democrats
November 11	12:00	1	state of Dayton peace talks
		2	Kohl and Verheugen comment on Bosnian peace talks
		3	Javier Solana to be named new NATO Secretary General
		5	trial of Nazi Erich Priebke begins in Italy
		6	German Supreme Court debates changes in political asylum law
		7	Scharping asks for vote of confidence
		9	debts of *Bundesländer* from the former East Germany
	12:30	1	state of Dayton peace talks
	13:00	3	three Polish secretaries of state step down
		4	Javier Solana to be named new NATO Secretary General
		5	comments of Däubler-Gmelin on asylum laws
		7	Kinkel defends policy of critical dialogue
	13:30	1	financing of fast Trans-European train networks
		5	new secretary of state for North Rhine Westphalia named
	14:00	1	state of Dayton peace talks
		8	juridical decision on pensions
	14:30	5	Rau will not leave politics
	15:00	1	state of Dayton peace talks
		3	Chinese dissident Wei Jingsheng imprisoned
		6	*Deutsche Bahn* may offer VISA card
	15:30	7	state of Dayton peace talks
	16:00	1	Daimler-Benz will restructure their aerospace subsidiary DASA
		8	German Supreme Court discusses asylum law
	17:00	2	Daimler-Benz will restructure their aerospace subsidiary DASA

Table 6.5. Overview of texts in DLF-RE. For each news item, its position in the news and its topic is given.

	FFH	HR
Day	News Items	News Items
June 16	2, 3, 4, 5	1, 2, 3, 4, 5, 6, 7
June 17	2, 3, 4, 5	1, 2, 3, 4, 5, 6
June 22	1, 2, 3, 5, 6	
June 24	1, 2, 4, 5	

Table 6.6. Overview of the FFH/HR-RE corpus

about news language which were derived from an analysis of the Stuttgart data.

The corpus consists of photocopies of the original manuscripts for 13 FFH shows and 13 HR shows. 9 pairs of shows were recorded at 15:55/16:00, the start of late afternoon prime time, and 4 pairs early in the morning at the start of breakfast time, 6:55/7:00. The shows were recorded on 13 separate days during June, July, and August 1992, with one pair of shows per day. FFH news typically last 3 minutes, HR news 4.5 minutes (Haaß 1994, page 26). Since the stations do not differ in the average number of stories, the time differences must be due to differences in story length and ultimately, in news selection and language. Only stories which were presented completely by the news reader himself were selected for analysis. Table 6.6 gives an overview of the selected texts. All texts come from June recordings.

Comparison: Although the German and the American English corpus both consist of radio news, they differ quite markedly in overall structure; one could even say that these texts belong to different genres (LaRoche 1991b). American and German radio stations differ markedly in their history and their organisation (Hoffmann-Riem 1985, Meyn 1999, Bliss 1992, Donsbach and Mathes 1999). To sum the main differences up in a sentence, German radio went from public to private, and American radio from private to public.

In Germany, almost all radio stations have a 3-5 minute news flash every hour; in the morning, most stations broadcast a review of the day's headlines at the half hour. The news is almost always followed by the weather forecast, and, for stations that broadcast traffic information, information about current traffic jams. Once a story is written, it can be reused in subsequent bulletins, if nothing significant has happened, or it is modified to report relevant changes as the day wears on. Because of this reuse, speakers in the German corpus sometimes read the same story several times during the day.

In America, not many stations adhere to such a strict schedule. In the U.S., news reports are the dominant form of radio news. They are designed to catch the listener's interest and come closer to the classical reportage than to condensed news flashes put together from agency reports (Stümpert 1991). In German radio, such reportages are usually left to special shows (Altrichter 1975), which can last several hours and interleave reports and interviews with music.

Because of these large differences in form and organisation, German and American radio news are more than culture-specific instantiations of the same genre—they constitute completely different genres in the sense of (Swales 1990), albeit with a similar communicative purpose. (For more on genres and text types, see the discussion in Appendix C.)

Code	Category
PRN	pronoun
DEF	definite article or possessive in determiner position or quantifier "all"
INDEF	indefinite article or cardinals in determiner position
NE	determinerless, head is a proper name
NA	no article, not a proper name

Table 6.7. Categories for the form of referring expressions used in AUDIX-4 and FFH/HR-RE

6.2.3 Annotations

For AUDIX-4 and FFH/HR-RE, the analysis is based on the news readers' manuscripts. Only the text itself is analysed, not additional material from reporters used to illustrate a news item. The DLF and WBUR texts are transcriptions of actual newscasts.

AUDIX-4 and FFH/HR-RE were labelled with boundaries of referring expressions and co-specification chains. For each referring expression, I determined whether it is a pronoun, a definite NP, an indefinite NP, a proper name, or neither (Table 6.7). I also calculated distance to last mention (first mention, last mention in same sentence, last mention in previous sentence, last mention before previous sentence).

The annotations of WBUR-LABNEWS and DLF-RE, on the other hand, are somewhat more complex. For each referring expression, I coded important aspects of its form, its syntactic function, its entity status, both in terms of the source-based scheme developed in Section 5.2.2 and in terms of the Givenness Hierarchy, and the semantics of the discourse entity that was specified. I also computed its position in the text. On the following pages, I will discuss the annotations in more detail.

Structure

The main unit of annotation is the word; sub-word units are not considered. Following common practice in computational linguistics, a word is an uninterrupted sequence of characters between two whitespaces. Referring expressions consist of one or more words. Parts of words do not constitute separate referring expressions, even if the word is hyphenated. This solution is not optimal from a linguistic point of view, but it guarantees reliable labels. Each word is assigned its position in the text: (paragraph or) sentence initial, (paragraph or) sentence final, and medial, that is, not at the boundary of a unit. These labels are computed directly from the annotations. If a word is at a boundary, we specify its position according to the largest unit which this boundary is associated with. For example, paragraph-final words are not additionally marked as sentence-final. Paragraphs were only annotated in WBUR-LABNEWS; they were present in AUDIX-4, but not labelled. For referring expressions, I also calculated depth of nesting. If a NP is the daughter of a VP, it has depth 0, if it is the daughter of a NP of depth n, it has depth $n + 1$.

Code	Category	Example
DEF	definite article	the German chancellor
INDEF	indefinite article	a spokesman for the American delegation
POSS	possessive pronoun / genitive noun	his decision to run / Gore's decision to run
CNT	numeral	two Democrats
QNT	quantifier	every Republican
NA	no determiner	Grand Old Party, Helmut Kohl

Table 6.8. Codes for Determiners in the Radio News Annotation

Syntax

The syntactic attributes describe relevant properties of the form of a referring expression and of its function in sentence. They were designed to be as succinct as possible. Since no parse was available, all information had to be hand-coded.

Five main form categories are distinguished: pronouns, including demonstrative and possessive pronouns, bare common nouns, bare proper names, prepositional phrases, and determiner phrases. For determiner phrases, I also coded the type of determiner (Table 6.8).

Four attributes code whether some frequent types of modifiers are present: AMOD for adjectives, PPMOD for prepositional phrases, NMOD for NPs and DPs, and RCMOD for relative clauses.

The syntactic function categories are summarised in Table 6.9. The class names are based on traditional terminology. The basic distinction is between grammatical *subjects*, obligatory arguments of the VP head (*objects*), and other NPs (*adjuncts*). Adjuncts are either optional VP arguments or NP arguments. The two subclasses can be distinguished by their level of embedding in other NPs: A VP argument is at level 0, an NP argument at level 1 or deeper. Most objects are either direct objects or prepositional objects; adjuncts tend to be genitive or prepositional adjuncts.

Both the three adjunct and the three object classes are motivated by surface properties of the NPs. Dative objects are only labelled in the German texts, because German distinguishes between accusative and dative objects inflectionally. English, on the other hand, uses a special preposition, "to", for dative objects, whose surface form becomes thus equivalent to that of PP objects.

In this study, I concentrate on choice of determiner, presence of modifiers, and syntactic function; word order correlates of entity status were left aside. This has two reasons: Firstly, throughout the thesis, I focus on the influence of entity status on the form of referring expressions, and this focus is maintained here. Secondly, a thorough analysis of word order would have required a full parse, which was not feasible given the time constraints.

Semantics

Semantic properties such as countability or genericity influence which determiner a referring expression will carry (Eisenberg 1994, Carlson 1977). Therefore, we should certainly include

Code	Constituent Type		Explanation
SJ	*subject*		obligatory argument of verb, nominative case
	object		subcategorised for by verb
DOBJ		direct	accusative case
GOBJ		genitive	dative case
POBJ		prepositional	prepositional phrase
	adjunct		non-obligatory VP argument or NP argument
PADJ		prepositional	in the form of a prepositional phrase
GADJ		genitive	genitive adjuncts
OADJ		other	NP/DP adjuncts, accusative or dative case

Table 6.9. Types of syntactic constituents. This syntactic classification relies on standard concepts of phrase structure grammar, but not on any specific theory.

some semantic information in the annotation scheme beyond that which is provided by part-of-speech tags (proper name vs. common noun) or syntactic form (bare NPs tend to refer to kinds).

From the wealth of different types of information one might want to encode, ranging from thematic roles to semantic features, three were selected: genericity, countability, and sortal class. Thematic roles were left out of the picture because any principled assignment of such roles, be it according to Jackendoff's (1990) Lexical Conceptual Semantics, according to Dik's (1989) Functional Grammar, or according to Halliday's (1994) Systemic Functional Linguistics, requires the analyst to determine the class of the verb first. This step introduces an additional source of errors.

Before we proceed with the annotation conventions, a note of caution: Many semanticists are bound to squirm at some of the annotation conventions proposed here. Countability (Carlson 1991), genericity and sortal classes are each of them fields with a long research tradition, and it is not possible to do that research justice in annotation conventions for corpora. I distilled the results of my main sources, (Krifka 1991, Carlson 1991) for countability and (Krifka et al. 1995) for genericity, into a set of easily applicable heuristics that were revised to cover difficult cases as I proceeded in my annotations. In her typological work, Behrens (1995, in preparation) has proposed an interesting classification scheme for referring expressions, which she has also applied to such thorny issues as genericity or the mass/count distinction. Since her category definitions are language-independent, annotations using that scheme will very likely be less circular than the combination of heuristics and semantic theory that I used here. If the scope of the present study is to be extended, the annotation scheme should clearly be revised to take her results into account.

Countability: The attribute CNT identifies six classes of nouns which have distinctive syntactic and morphological properties in German and English: proper names (PN), collective nouns (COLL), and mass nouns (MASS). NPs which refer to concepts and which are not countable are labelled (MABS), and NPs which belong to neither of these categories are labelled Y (for countable: yes). The attribute is not assigned based on the semantics of the discourse entity

Code	Category	Examples
PN	proper name	Michael Dukakis, the Safe Roads Act Hampton County jail, 1999, Washington
MABS	uncountable concepts	availability, hope (general feeling)
MASS	mass nouns	water
COLL	collectives	police, board of directors
Y	countable	schools
NONE	not applicable	he, she, it

Table 6.10. Countability categories

itself. Rather, it sits on a fine line between syntax and semantics—on one hand, it encodes a morpho-syntactic distinction, pluralisability, on the other, semantic concepts that have been the subject of vociferous debates.

Let us begin with the purely semantically motivated distinctions, collectives and mass nouns. Both types of nouns do not distinguish singular and plural forms and cannot be combined with numerals. Mass nouns can be distinguished from collective nouns in that mass nouns refer to homogeneous entities without natural partitions (Krifka 1991). Table 6.10 shows some examples. In the radio news texts, mass nouns are very rare, collective nouns as well. The semantically interesting distinctions are thus almost irrelevant for our statistical analysis—not enough reports about police violence or water shortages, I'm afraid.

The category "proper name" covers not only names of persons, places, laws, buildings, or companies, but also temporal expressions that name specific dates and times. For example, in the phrase "On Monday, September 25, she handed in her thesis.", the temporal expression "Monday, September 25" is treated as a name, while in the phrase "On Wednesdays, she usually goes for a swim with Gerd.", "Wednesdays" is treated as a type-identifiable expression. Time spans are labelled as MABS.

For referring expressions that denote persons, the category PN is only assigned when there is an extensional reference to that person using the name. NPs that refer to a person by their function are treated as countable, even if the name of that person is specified in a modifier. In this interpretation, the proper name modifier narrows down the interpretation of the head NP to a single individual, that designated by the proper name.

Plurals are labelled as countable according to the classification of the corresponding singular. Since the attribute does not really make sense for pronouns, it has the default specification "none" there.

Genericity: Generic passages are quite rare in the data. Most of these passages describe consequences that legal decisions will have for the listeners. They are part of so-called service news items (Zehrt 1996), which are rare on DLF, a station that tends to stick to hard news in its on-the-hour news flashes. Hence, most generic sentences fall in the category of lexical habituals (Krifka et al. 1995). An example follows:

(6.1) Zudem sollen Überstunden nur noch in der Freizeit abgegolten und die Lohnnebenkosten gesenkt werden.

Code	Category	Examples
PER	persons and non-institutionalised groups of persons	drunken drivers, Michael Dukakis, three single mothers
PLACE	places	Massachusetts
TIME	time/date	last year
INST	institutions, political entities, companies, institutionalised groups of persons	the Board of Representatives, Massachusetts Supreme Court, Massachusetts
THING	physical thing, sums of money	car, road
EVENT	event and type of event	subscription, application, negotiation, computerised phone calls
ABSTRACT	other abstract concepts	birth control, Safe Roads Act

Table 6.11. Sortal classes with the corresponding attribute value and an example

This sentence concerns a job creation programme. *Überstunden* (overtime) is certainly generic, because the proposal refers to overtime in general. The NP "*Lohnnebenkosten*" (additional salary costs) is problematic: either it refers to *Lohnnebenkosten* in general, or to a specific economic variable.

Sortal Class: The sortal classes for the radio news texts are listed in Table 6.11 together with some examples. When assigning a class to a referring expression, the classes are considered in the order in which they appear in the table until a class has been found to which the discourse entity can be assigned satisfactorily. The last class, "abstract concepts", is a sort of garbage class. The categories chosen differ somewhat from those that were used for BROWN-COSPEC (Appendix C). States and processes, which are usually realised by verb phrases, have not been assigned separate categories, because they occur rarely as nominalisations. On the other hand, institutions have been singled out because they occur quite frequently in the texts.

Entity Status

To label the corpora with all taxonomies of entity status ever designed in order to compare their empirical coverage would have been pointless, because many taxonomies are but reduced versions of fuller ones. For AUDIX-4 and FFH/HR-RE, I operationalised entity status as distance from last mention, following the strategy of e.g. (Ariel 1990, Givón 1992). WBUR-LABNEWS and DLF-RE, the corpora for which I have access to the speech files, were additionally labelled with two taxonomies: the source-based scheme introduced in Section 5.2.2, and the Givenness Hierarchy (Gundel et al. 1993, c.f. also Table 4.5). The Givenness Hierarchy labels were placed according to the original publication; as far as I know, no annotation manuals or training material are available. This makes annotating new material difficult because in annotation, many minor issues crop up that are best settled by an extensive manual. That the annotations

have not been validated by another annotator means that they represent my judgements about the management of discourse entities in these texts.

I have not investigated issues of topicality further on this data. The main reason is the German data: many stories are sequences of sentences with almost no overt contextual links. Not many sentences have topics, and thematic progression tends to be unordered. The problems are discussed in more detail in Section 6.4.1. Neither did I analyse my data in terms of Centering Theory. As we have seen in Section 4.3.2, there are many competing takes on ordering the list of forward-looking centres and the proper definition of units. A proper, principled analysis should compare at least two or three of these alternatives in depth, and that was not possible for reasons of time. Furthermore, in news discourse, referential continuity does not appear to be very important for establishing coherence. A prime example is the Gemayel text, a perfectly normal news report. Hence, we would expect any Centering analysis to yield plenty of rough shifts, and that is exactly the result a preliminary analysis of some discourses using the standard algorithm (Brennan et al. 1987) gave.

6.3 Quantitative Analysis

I now turn to the first part of my analyses of entity status in news discourse, the quantitative analyses of the corpora described in Section 6.2. To begin with, Section 6.3.1 surveys the distribution of referring expressions and co-specification sequences in the four corpora. Next, Section 6.3.2 deals with influences on the form of referring expressions other than entity status, such as the semantics of the discourse entity. After this groundwork, we can set about quantifying how the status of a discourse entity influences the way in which it will be mentioned. This analysis will focus on three questions:

1. How are new discourse entities introduced? Is there a difference between *tracking* referents, those that will be mentioned frequently, and *deadend* referents, those that will not be mentioned but once? (Section 6.3.3)

2. How are old discourse entities accessed? (Section 6.3.4)

3. Can we find correlates of the different taxonomies of entity status described in Section 5.2? Do some taxonomies have more clear-cut correlations than others? (Section 6.3.5)

Note on Percentages and Accuracy: In order to eliminate mistakes as far as possible, I checked the annotations of the corpora myself repeatedly, searched automatically for inconsistent feature combinations, and took care to add the annotations in layers, following the recommendation of (Hirschman et al. 1998). First I labelled the boundaries of referring expressions, next, I established co-specification sequences, and finally, I inserted the attributes of the referring expressions. In each step, I caught a few earlier mistakes. Nevertheless, since it was not possible to have another annotator cross-check the annotations of all four corpora, I must allow for a certain margin of error in my statistical results. To be on the safe side, all differences between percentages which are smaller than 1% are bound to be variations within the margin of annotator error, and contingency table cells with less than 5 items are treated with due caution. To avoid such sparse cells, I will also collapse categories which, though semantically well

	German			American	
	DLF	FFR	HR	WBUR	AUDIX
# stories	31	17	13	4	4
total story length (words/clauses)	2791 / 149	1073 / 70	1047 / 63	2112 / 112	1927 / 97
avg. story length (words/clauses)	90 / 5	63 / 4	81 / 6	528 / 28	482 / 24
# discourse entities	646	252	214	323	326
% co-specification sequences	86 (13%)	45 (18%)	52 (**24%**)	113 (**35%**)	46 (14%)
# referring expressions	787	302	293	547	439
avg. # referring expressions	25	18	23	137	110
discourse-old r.e.	141 (18%)	50 (17%)	79 (25%)	232 (**42%**)	113 (26%)
pronouns	49 (6%)	15 (5%)	20 (7%)	88 (**16%**)	28 (6%)
definites	409 (52%)	142 (47%)	141 (48%)	175 (**32%**)	131 (**30%**)
indefinites	87 (11%)	41 (14%)	32 (11%)	38 (**7%**)	75 (17%)
proper names	120 (23%)	57 (19%)	62 (21%)	66 (**12%**)	98 (22%)
bare NPs	122 (**8%**)	47 (16%)	38 (13%)	180 (33%)	107 (24%)

Table 6.12. Distribution of referring expressions in the four corpora AUDIX-4, DLF-RE, WBUR-LABNEWS, and FFH/HR-RE. Percentages in brackets; significantly high or low percentages (Fisher test, $p<0.01$) are in bold face.

motivated, rarely occur in the data. This concerns in particular the determiner categories QNT (quantifier) and CARD (cardinal), which occurred less than ten times in each of the corpora.

All percentages will be rounded to the next full integer. For this reason, some percentages may not add up to 100. I chose this rather coarse rounding because the corpora are so small that one single change of annotation can change percentages by as much as 0.34 percentage points.

6.3.1 Baseline I: Differences Between the Corpora

Table 6.12 shows that the distribution of referring expressions in the four corpora differs greatly. One fundamental difference is due to the language: In German, bare NPs are permitted in fewer contexts than in English. The comparison is also made more difficult by the fact that, as Table 6.15 shows, the German data contains very few generics.

In both languages, possessives and bare NPs frequently co-occur with generic discourse entities. The generic pronouns in WBUR-LABNEWS come from longer passages which describe the consequences of new laws. In one passage of text R, the generic pronoun is the second person "you": the author directly addresses the listener to tell him that if he is caught driving drunk, he faces certain inconvenient consequences, although she is actually talking about potentially drunk Massachusetts citizens in general.

In the German corpora and in the AUDIX-4 agency copy, we find comparatively many first mentions, corroborating the results of Biber (1992), which were summarised in Section 5.1.1. Only WBUR-LABNEWS departs from that pattern. This is also the corpus with the highest percentage of pronouns, and the lowest percentage of indefinites and proper names. The reason for this radical difference is simple: the WBUR-LABNEWS texts are reports, they develop a

Corpus	subject	object	prep. adjunct	other adjuncts
WBUR-LABNEWS	39 %	**30** %	17 %	13 %
DLF-RE	33 %	19 %	**34** %	14%

Table 6.13. Distribution of syntactic functions. (Percentages are rounded and therefore do not add up to 100.)

Corpus	person	inst.	place	time	thing	abstract	event
WBUR-LABNEWS	**34** %	13 %	3 %	3 %	**7** %	**38** %	3 %
DLF-RE	24 %	19 %	9 %	5 %	3 %	23 %	**16** %

Table 6.14. Distribution of sortal classes.

coherent story and give necessary background information. The other corpora focus on getting the main news message across. This is also true for the longer AUDIX-4 agency reports. Upon reading them, it becomes clear that they are mainly a collection of relevant material which still needs to be copy-edited.

Looking at the distribution of referring expressions over grammatical roles, we observe that the frequency of adjuncts is far higher in DLF-RE (Table 6.13). This effect can be traced to the genre "news flash". As much information as possible has to be crammed into a few sentences. This information should answer the classical questions that a news story is supposed to answer: Who did What to Whom Where and When? Where and When are prime candidates for adjuncts. Modifiers of referring expressions whose function it is to help listeners identify the referent quickly and efficiently.

Furthermore, there are marked differences in the distribution of semantic classes (Table 6.14). While about a third of all referring expressions in the WBUR texts refer to people, and a third to abstract concepts, this drops to a fourth each in the German texts. In contrast to WBUR-LABNEWS, NPs in DLF-RE refer much more frequently to events. There are several reasons for this effect. First editors tend to report events not in VPs, but in NPs, with the event expressed by a nominalisation and a semantically bleached main verb such as "stattfinden" (*took place*). Second, reporting events as NPs makes it easy to mention several events in the same sentence, events that took place before that described in the main VP or events that stand in a causal relationship with it. Lastly, there is a markup-specific reason: nominalisations in German are often expressed by gerund constructions in English. Since gerunds are verb forms, they do not count as referring expressions.

6.3.2 Baseline II: Semantic Influences

Entity status is certainly not the only influence on the choice of article. Other important influences are countability and genericity, and to a certain extent also sortal class. For example, generic statements about kinds frequently involve bare plurals (Carlson 1977). In this section, I describe the relevant patterns that surface in the data, before I discuss how entity status may modify them. If not stated otherwise, all significant associations were found by a Fisher test, significance level $p<0.005$. The significance level was chosen because the analysis involves

WBUR-LABNEWS							
	pronoun	name	def.	poss.	indef.	bare NP	total
generic	36%	0%	26%	49%	39%	52%	194 (35%)
specific	64%	100%	74%	51%	61%	48%	353 (65%)

DLF-RE							
	pronoun	name	def.	poss.	indef.	bare NP	total
generic	10%	0%	3%	10%	7%	15%	43 (5 %)
specific	90%	100%	97%	90%	92%	85%	745 (95%)

Table 6.15. Effect of Genericity on Form of Referring Expressions. Percentages are based on absolute frequency of pronouns / determiner type. Last column: absolute frequency of generic/specific in corpus

many significance tests, some of which might yield spurious positives.

Sortal Class: Which modifiers we can expect to find in a referring expressions depends on the sortal class to which it belongs.

In WBUR-LABNEWS, abstract nouns tend to have more PP and adjective modifiers. Institutions etc., on the other hand, show more NP modifiers. One reason for this is that in phrases such as "Boston University School of Public Health" (text R, WBUR-LABNEWS) or "Massachusetts Supreme Court" (text J, WBUR-LABNEWS), I analysed the NP modifiers "Boston University" and "Massachusetts" as a separate referring expression. Incidentally, "Massachusetts" is part of a co-specification sequence that stretches through the whole text. Persons are less likely to be referred to with a definite NP, and less likely to be described with adjectives. In fact, most persons in these texts are referred to either by their name (and social function), or by a pronoun. In DLF-RE, place names tend not to occur with definite determiners (29.87% vs. 52% overall)

In both corpora, events are significantly more likely to come with the indefinite article. In DLF-RE, 18.70% of event NPs have an indefinite article, but only 6.98% of all referring expressions in the corpus have it. In WBUR-LABNEWS, the numbers are 26.67% and 5.67%, respectively. For Event NPs of that type are frequently used when the author gives some background information.

Genericity: The definite article is significantly less frequent with generic referents than with specific ones. This holds for both corpora (Table 6.15). However, generic referents and sentences with generic readings are quite rare in DLF-RE, because these texts report on specific events. Genericity becomes relevant only when reporting the reasons politicians have given for taking certain measures, or when explaining the consequences of political events for the general public. Generic referents thus occur within prototypical background information, which can be cut if necessary.

WBUR-LABNEWS					
	def.	poss.	indef.	bare NP	total
named	6%	0%	0%	0%	79 (14%)
countable	86%	89%	97%	78%	412 (75%)
coll. + mass	2%	0%	0%	2%	8 (1%)
uncountable	6%	11%	3%	20%	49 (9%)

DLF-RE					
	def.	poss.	indef.	bare NP	total
named	32%	10%	0%	0%	248 (31%)
countable	49%	48%	84%	70%	407 (52%)
coll. + mass	2%	0%	2%	2%	13 (2%)
uncountable	17%	42%	14%	28%	122 (15%)

Table 6.16. Distribution of determiners for countable vs. mass nouns/collectives vs. uncountable head nouns of referring expressions. For a definition of the countability values, see Table 6.10.

Countability: Mass nouns such as "water", "blood" usually do not take an article. However, both mass nouns and collectives are rather rare in these corpora. Much more frequent are abstract nouns which occur only in either singular or plural. They account for 8.8% of all referring expressions in WBUR-LABNEWS and for 9.6% of all referring expressions in DLF-RE. Such nouns are often article-less in the American English corpus, but not in the German corpus, where they tend to be associated with the definite article. This might be a language-specific difference in article use, but since the subject matter of the two corpora also differs considerably, it could also be the case that different types of abstract nouns occur in each corpus. DLF-RE has considerably more proper names than WBUR-LABNEWS (27.7% vs. 13.5%). The definite article is significantly less frequent with proper names; still, it occurs for some institutions and for persons who are introduced both with their name and their function.

Summary: Since the texts barely contained any natural kinds, classic correlates of genericity were hard to measure. However, the definite article did occur significantly less frequently with generics in both corpora. There were some effects of sortal class. The effects of countability were less clear-cut, partly because classic mass nouns and collective nouns were rather rare, partly because assigning all abstract nouns that cannot be declined for number to one class could potentially obscure distinctions which are important for the distribution of articles.

We have seen that the semantics of a discourse entity plays an important role in determining the form of a referring expression. In the following sections, we will look at the effect of entity status, focusing on the management of discourse entities. In Section 6.3.3, we will see how the particular communication situation in radio news affects the way new entities are introduced, and in Section 6.3.4, I examine how these entities are maintained and accessed. Finally, I ask whether any of the taxonomies of entity status defined in Section 5.2.2 can predict aspects of the form of referring expressions in radio news discourse.

	WBUR-LABNEWS	AUDIX-4	DLF-RE	FFH/HR-RE
Definites	32%	**26**%	52%	49%
Indefinites	**9**%	**23**%	**9**%	**15**%
Proper Names	**9**%	20%	14%	**17**%
Bare NP	**42**%	**30**%	**18**%	17%

Table 6.17. Percentage of first mentions realised as definites / indefinites / proper names / bare NPs. **bold**: significant association between category and discourse-new

6.3.3 Introducing New Entities

From our discussion of the genre of radio news in Section 6.1, it follows that most first mentions should be definites. The rationale for this is simple: Definites tend to be familiar and uniquely identifiable (leaving the chicken-and-egg problem of which of the two is primary aside for a moment). News items tend to be about familiar news actors and nations. In order to give the audience a fair chance of understanding what the news item is about before the text is over, first mentions should also be uniquely identifiable. This requirement is less strict for the longer American texts that can spend more time elaborating on background information. A cursory analysis of the texts already confirms this hypothesis: All texts show a marked tendency to introduce even relatively well-known referents by some property which makes them uniquely identifiable, usually their public function or their position in the organisation that is being talked about. For example, Michael Dukakis is referred to as "Governor" when he is first mentioned in text J, WBUR-LABNEWS, and then-chancellor Helmut Kohl, a more than familiar figure to most Germans, is always introduced as "Bundeskanzler Kohl" (chancellor Kohl) when he is mentioned in the German news items. This description corresponds to a proper name (Kohl) plus an attributive noun which helps listeners identify Kohl by his function. Such first mentions instantly call up the scripts associated with these people in their official function as Governor or Chancellor. As we can see in the analyses of section 6.4.2, this cue can be adapted depending on the news item. The quantitative results summarised in Table 6.17 present a somewhat more complex picture. Although definites (including NPs with a possessive in determiner position) account for roughly half of all first mentions in the German corpora, that percentage dwindles to a third (WBUR-LABNEWS) or a fourth (AUDIX-4) in the English corpora. Significance tests (Fisher test, $p<0.005$) indicate that there is no statistically significant association between definite descriptions and discourse-new entities, except for AUDIX-4, the (comparatively) unedited agency copy. Rather, definite descriptions tend to be distributed equally across first and subsequent mentions. In the DLF-RE-data, definites appear to be reserved for co-specification sequences: first mentions that start a new sequence are significantly more likely to be definites.

In general, the indefinite is the only really reliable indicator that a referring expression evokes a new discourse entity is the indefinite article. However, since it is less important that deadend discourse entities be identifiable, we would expect that the indefinite article is more strongly associated with them than with first mentions of tracking entities. Table 6.18 confirms that hypothesis: the indefinite mostly occurs with deadend entities, entities that are never accessed again and were merely introduced to supply background knowledge. In AUDIX-4, FFH/HR-RE and DLF-RE, bare NPs perform a function that is similar to that of indefinites:

	def.	(poss.)	indef.	name	bare NP	pronoun	total	
DLF-RE								
first mentions	84%	(100%)	100%	75%	97%	6%	646	(82%)
deadend	71%	(97%)	92%	65%	91%	2%	560	(71%)
tracking	14%	(3%)	8%	10%	6%	4%	86	(11%)
hearer new	63%	(93%)	95%	81%	23%	6%	471	(60%)
new anchored	8%	(59%)	2%	0	0	0%	37	(5%)
FFH/HR-RE								
first mentions	81%		96%	100%	68%	0%	466	(78%)
deadend	64%		81%	93%	41%	0%	379	(64%)
tracking	17%		15%	7%	27%	0%	87	(14%)
WBUR-LABNEWS								
first mentions	63%	(91%)	95%	42%	76%	13%	322	(59%)
deadend	56%	(69%)	63%	21%	54%	3%	212	(39%)
tracking	7%	(22%)	32%	21%	22%	10%	110	(20%)
hearer new	42%	(63%)	55%	17%	22%	5%	132	(24%)
new anchored	25%	(60%)	29%	8%	12%	3%	82	(15%)
AUDIX-4								
first mentions	66%		99%	66%	93%	7%	326	(74%)
deadend	53%		88%	47%	91%	7%	280	(64%)
tracking	13%		11%	19%	2%	0%	46	(10%)

Table 6.18. Forms of first mentions. Percentages are based on the absolute frequency of the types of referring expression in the corpora.

they are used for entities that are referred to but once. This tendency is less pronounced in WBUR-LABNEWS: most of the texts are about specific issues, and key concepts that relate to these issues, such as "birth control" in text T, are always determinerless. We do not get that effect in AUDIX-4 because the texts I selected from that corpus are classical news items, reports about events and actions.

For DLF-RE and WBUR-LABNEWS, which I coded using the full source-based scheme, Table 6.18 shows how hearer new and brand-new anchored discourse entities are realised. In both corpora, bare NPs tend to be hearer-old concepts. Names are much more often hearer-new in DLF-RE. This is partly due to my restrictive assumptions about the audience (John Doe, c.f. page 137), partly to the fact that the news items report events from all over the world. In DLF-RE, only definites show no significant correlation with hearer old/new. The picture is different for WBUR-LABNEWS: in that corpus, only indefinites and possessives are used more frequently with hearer-new discourse entities than with hearer-old ones.

WBUR-LABNEWS				
	adjective	NP	PP	relative clause
total	12%	12%	18%	5%
first mention	24%	15%	19%	8%
hearer new	24%	25%	33%	12%

DLF-RE				
	adjective	NP	PP	relative clause
total	21%	19%	13%	1%
first mention	23%	22%	15%	1%
hearer new	26%	23%	19%	1%

Table 6.19. Entity status and modifier use in radio news. total: baseline; first mention: discourse new; hearer new: unknown to hearer

When new discourse entities are anchored, they mostly occur with a definite or indefinite article. In WBUR-LABNEWS, some names and bare NPs are anchored as well. The numbers might suggest that anchoring is not very frequent in the corpora. In particular, the editors of DLF-RE fail to anchor discourse-new entities to discourse-old ones. But this analysis overlooks that discourse-new entities can also be anchored in the hearer's world knowledge.[5] All news actors are introduced with modifiers that describe their (mostly political) function. For example, the name of the Danish Foreign Secretary eludes me at the moment, but I know that there must be such a person in Denmark, and when I encounter the name "Uffe Ellemann-Jensen" in DLF-RE as a modifier of the NP "Danish Foreign Secretary", I can immediately place him. Table 6.19 documents that in both WBUR-LABNEWS and DLF-RE, modifiers are used more frequently for first mentions than for subsequent mentions. There is a good reason for this pattern: if a new discourse referent is not hearer-old, and if it is very difficult or impossible to introduce it via bridging, it makes sense to introduce modifiers such as attributive NPs, adjectives, PPs, and relative clauses, since they can carry information which will help the addressee identify the discourse entity that the referring expression introduces.

6.3.4 Accessing Old Entities

We expect to find three patterns in our data:

1. The distance between anaphoric definite referring expressions and their antecedents is larger than the distance between anaphoric pronouns and their antecedents. In particular, pronouns are preferred intra-sententially and inter-sententially if the antecedent is in the previous sentence, else, we get definite descriptions. This is the standard pattern (c.f. e.g. McCoy and Strube 1999).

[5] I use the term "anchored" because it is a good metaphor; but it would also be reasonable to reserve anchoring for the more specific operation of "anchoring in discourse-old information".

2. Persons are more likely to occur in co-specification sequences, more salient, and hence more likely to be referred to by a pronoun than other sortal classes. This pattern is suggested by the news factor PERSONALISATION —events tend to be reported as the actions of persons—and ATTRIBUTION—it is important to have élite sources, and to name them.

3. Discourse-old entities will tend to surface in subject position and as (possessive) genitive adjuncts; they anchor the new information in a sentence to the preceding discourse.

We will deal with each of these three hypotheses in turn.

Distance to Last Mention: Our first hypothesis is partially confirmed: pronouns predominate only when the antecedent is in the same unit; else, definites, names, and bare NPs are used, even when the antecedent is in the previous sentence. Table 6.20 presents detailed results for definites, proper names, and pronouns. But contrary to what we find in the BROWN-COSPEC-corpus (c.f. Table 7.7), pronouns are rarely used for inter-sentential anaphora—this is the domain of definites and proper names. The pattern reflects a tendency that we already encountered on page 132 ff.: Anaphoric definites convey new information about the discourse entity they specify. We thus get the curious case of an expression that refers to something old—the discourse-old discourse entity—with "new" information. This shows very clearly that, just as e.g. Lambrecht (1994) argued, Givenness should be separated into two dimensions, the givenness of discourse entities, which can be separated in turn into identifiability and activation, and the givenness of information. I will pursue that point further in my concluding arguments in Section 8.2.

Table 6.20 also shows clear differences between the genres. The edited reports of WBUR-LABNEWS come closest to the patterns of BROWN-COSPEC as documented in Appendix 7.1.2. AUDIX-4, the agency copy, already deviates from this pattern. Although every text is about one main topic and although the texts are as long as those from WBUR-LABNEWS, the texts have less pronouns. On reading them, they appear to be less coherent. The reason for this is simple: The texts are intended as raw material for editors; too much polishing will be a waste of time if your audience will only use arbitrary snippets of what you have written, anyway. Most of the names refer to persons and places (c.f. Table 6.21).

Sortal Class: The data in Table 6.21 largely confirm our second hypothesis: Pronouns account for more than half of all subsequent mentions of people and physical objects. This tendency is far more pronounced in DLF-RE than in WBUR-LABNEWS: Not only are but 4 of all 68 references to abstract objects discourse old, but none of them is realised by a pronoun. Discourse-old persons, on the other hand, are pronominalised as frequently as in WBUR-LABNEWS. Events, times, abstract concepts, and institutions are mostly referred back to by definites. The results on BROWN-COSPEC suggest that the same patterns hold in academic prose, although these texts can be said to be "about" abstract concepts, while news texts are arguably "about" people and events.

Syntax: Again, the data confirm our hypothesis. But there is no syntactic position on which either first mentions or subsequent mentions have a lock (Table 6.22). In both texts, first mentions dominate in object position and as prepositional adjuncts, while subsequent mentions are more likely to be subjects and genitive adjuncts. But the first mentions in non-subject position

	WBUR-LABNEWS				AUDIX-4			
	pronoun	definite	name	total	pronoun	definite	name	total
same unit	72%	18%	4%	67 (29%)	83%	11%	9%	18 (16%)
previous unit	35%	15%	36%	75 (32%)	18%	34%	39%	38 (34%)
before prev. unit	3%	29%	29%	87 (38%)	7%	52%	32%	57 (50%)

	DLF-RE				FFH/HR-RE			
	pron.	definite	name	total	pron.	definite	name.	total
same unit	88%	6%	6%	32 (23%)	68%	9%	24%	34 (26%)
previous unit	28%	48%	22%	60 (43%)	14%	50%	32%	66 (51%)
before prev. unit	2%	67%	24%	49 (35%)	10%	59%	31%	29 (22%)

Table 6.20. Distance to last mention for determiner types and modifiers. Percentages are relative to the row totals for each corpus

WBUR-LABNEWS							
	person	inst	place	time	thing	abstract	event
% discourse-old	53%	53%	41%	7%	37%	33%	7%
% pronouns	48%	11%	0%	0%	53%	27%	0%
% definites	15%	35%	0%	100%	20%	42%	100%
% names	23%	16%	86%	0%	0%	6%	0%

DLF-RE							
	person	inst	place	time	thing	abstract	event
% discourse-old	41%	18%	9%	5%	8%	6%	12%
% pronouns	48%	22%	0%	0%	50%	0%	27%
% definites	26%	68%	29%	100%	50%	72%	73%
% names	26%	11%	71%	0%	0%	0%	0%

Table 6.21. Effect of sortal class on the form of discourse-old referring expression. Percentage of mentions which are discourse-old, and the percentage of discourse-old mentions which are realised by a pronoun, a definite NP, or a name.

WBUR-LABNEWS						
	subject	obj.	prep. obj.	prep. adj.	gen. adj.	total
first mention	48%	71%	83%	73%	25%	59%
subsequent mention	52%	29%	17%	27%	75%	41%

DLF-RE						
	subject	obj.	prep. obj.	prep. adj.	gen. adj.	total
first mention	71%	95%	88%	93%	32%	82%
subsequent mention	29%	5%	13%	7%	68%	18%

Table 6.22. Distribution of first vs. subsequent mentions across syntactic positions

WBUR-LABNEWS					
	subject	obj.	prep. obj.	prep. adj.	gen. adj.
tracking	54%	29%	28%	17%	38%
deadend	46%	71%	72%	83%	63%

DLF-RE					
	subject	obj.	prep. obj.	prep. adj.	gen. adj.
tracking	28%	10%	4%	8%	8%
deadend	72%	90%	96%	92%	92%

Table 6.23. Distribution of first mentions across syntactic positions: deadend (mentioned only once) vs. tracking (start of a co-specification sequence)

are rarely taken up again anaphorically. Table 6.23 shows that first mentions in subject position are more likely to start a new co-specification sequence than those in other positions. In fact 57% of all co-specification sequences in DLF-RE begin with a referring expression in subject position; for WBUR-LABNEWS, this figure is 51%. The next most popular positions for starting a new sequence are prepositional objects for DLF-RE (24%) and direct and prepositional objects for WBUR-LABNEWS (each 16%).

6.3.5 How Useful are Detailed Taxonomies?

In Sections 6.3.3 and 6.3.4, we have relied mainly on aspects of entity status which are easy to operationalise: first mention vs. subsequent mention, distance to last mention, and anchoring. In this section, we explore the association between more complex measures of entity status and the form of a referring expression. From now on, I will focus on WBUR-LABNEWS and DLF-RE, the two corpora which were annotated in greater detail because full prosodic labels are available for both.

The central question is: To what extent can the form of a referring expression be predicted from the status of the underlying discourse entity? To make the statistical analysis easier, I collated the different variables used so far to characterise the form of a referring expression into

	TI	REF	UI	FA	AC	IF
WBUR-LABNEWS	19%	5%	26%	17%	13%	20%
DLF-RE	13%	12%	33%	24%	12%	6%

Table 6.24. Frequency of Givenness Hierarchy categories (c.f. Table 4.5)

STAT3	new	med									old	
STAT4	BN		AC								U	A
SOURCE	BU	BA	SIT	FRAME	PART	VAL	ISA	SET	EVENT	INF	U	AC
WBUR-LABNEWS												
	9.5%	15.0%	4%	10.0%	0.5%	0.0%	0.5%	4.2%	1%	0.5%	13%	41%
DLF-RE												
	35%	5%	3%	11%	0.5%	1%	0.5%	0.0%	1%	3%	22%	18%

Table 6.25. Frequency of categories from source-based scheme, c.f. Section 5.2.2

one meta variable FORM with five values: BARE for bare NP, PRO for a pronoun, DEF for the definite article, INDEF for the indefinite article, and PN for all proper names.

This section is structured as follows: Finally, I investigate whether one of these taxonomies of entity status, together with some information about the semantics of the discourse entity, is sufficient to predict the form of referring expressions.

First, let us examine how the distribution of the the categories of the more fine-grained taxonomies differ in the two corpora. With respect to Givenness as measured by the Givenness Hierarchy, the two corpora WBUR-LABNEWS and DLF-RE differ in two main aspects: DLF-RE has more referring expressions which specify uniquely identifiable or familiar discourse entities, while in WBUR-LABNEWS, more discourse entities are in focus (c.f. Table 6.24). Again, this reflects the fact that the two corpora come from different genres: on one hand, we have long reports, or *Berichte*, in Bucher's (1986) terms, on the other hand, pure information (*Meldungen*). This difference becomes even clearer when we look at the frequency of the categories in the source-based scheme (Table 6.25): DLF-RE contains far more references to (as yet) unused entities and to brand-new unanchored ones than WBUR-LABNEWS. The detailed subcategorisation of inferrables appears unnecessary, at least for news texts which cover a broad range of topics: In both corpora, frame-based inferences dominate by far.

Before we move on to significance testing, let us first look at some quantitative measures of the strength of the association between different taxonomies of entity status and FORM. For this purpose, we will use $\lambda_{\max}$ and Goodman and Kruskal's τ, which are described in Appendix B.4. The relevant results are summarised on Table 6.26.

The τ values (Eq. B.14) allow us to compare how much of the variation in form is explained by entity status. Although none of the taxonomies is perfect, the Givenness Hierarchy emerges as a clear winner. In both corpora, its τ-value is largest. In WBUR-LABNEWS, SOURCE and KDIST explain similar amounts of variation—even though SOURCE is far more complex. For DLF-RE, the performance of KDIST is roughly equivalent to that of STAT4, the reduced version

		SOURCE	STAT4	STAT3	DISC	HEAR	KDIST	GHIER
WBUR-LABNEWS	$\lambda_{\max}$	0.22	0.20	0.22	0.39	0.26	0.30	0.28
		54%	15%	11%	10%	6%	22%	36%
	τ	0.103	0.068	0.055	0.016	0.040	0.114	0.169
DLF-RE	$\lambda_{\max}$	0.20	0.25	0.24	0.38	0.37	0.37	0.36
		49%	18%	12%	9%	9%	28%	45%
	τ	0.115	0.071	0.050	0.035	0.044	0.077	0.236

Table 6.26. Association between taxonomies and form of referring expressions. For $\lambda_{\max}$, I both give the absolute value and the value in percent of the highest possible λ for that combination of categories

of SOURCE. Of the two dichotomies HEAR and DISC, HEAR has the greater explanatory power.

$\lambda_{\max}$ measures to what degree we could claim that entity status is the *cause* of the particular form observed. The connections are strongest for the two most detailed taxonomies, SOURCE and GHIER. They are followed by KDIST, a variable which, as I have argued in section 5.3, measures structural entity status. The other taxonomies are far less powerful.

τ and $\lambda_{\max}$ provide two different perspectives on the strength of the association between form of referring expression and entity status. Generalised linear models offer yet another test of predictive power: How well does entity status predict specific aspects of form, and can additional information about the semantics of the discourse entity improve that prediction? The methodology is similar to that in Section 7.2.1: Starting with a pool of predictor features, in each step, we add that predictor to the model which improves the model most. The quality of the model is measured by AIC (An Information Criterion), which takes into account both model size and model fit; the smaller the AIC, the better. We keep adding the predictor which reduces AIC by the largest amount until either no predictor can improve AIC or all predictors have been used. This procedure is called *stepwise forward selection*. For more on logistic regression, consult Appendix B. The basic set of predictors used here consists of the three semantic features SEM, CNT, and GEN plus distance from last mention (KDIST). For each of the six taxonomies DISC, HEAR, STAT3, STAT4, GHIER, and SOURCE, I added the taxonomy to the base set and ran stepwise forward selection. The methodology is applied to three tasks: deciding whether a referring expression should be realised as a pronoun, deciding whether it should be realised with a definite determiner, and finally, deciding whether it should be realised as a bare NP.

We begin with pronominalisation. As we have seen in Chapter 4, most arguments in the debates about givenness and most applications in anaphora resolution and generation focus on these forms. I will explore constraints on pronominalisation in more detail in Chapter 7. In that study, entity status is operationalised as distance from last mention, and sortal class is the only semantic information about discourse entities that was coded in the corpus. Both decisions were motivated by methodological considerations: distance is the most reliable measure of entity status there is, and sortal class can be annotated reliably if an external knowledge source is used that supplies the relevant ontological information. On the radio news data, I examine whether these decisions can also be justified empirically. For both WBUR-LABNEWS and DLF-RE, only GHIER and KDIST were included in any of the final model formulae. DISC,

WBUR-LABNEWS							
	model		**AIC after including feature**				
	formula	AIC	KDIST	tax.	CNT	SEM	GEN
default	AIC = 484.59						
KDIST only	KDIST + CNT + SEM	292.82	306.43	n.a.	296.52	*292.82*	no
KDIST and ...							
GHIER	GHIER + KDIST + CNT + SEM	266.85	276.83	303.83	269.02	*266.85*	no

DLF-RE							
	model		**AIC after including feature**				
	formula	AIC	KDIST	tax.	CNT	SEM	GEN
default	AIC = 369.1						
KDIST only	KDIST + CNT + SEM + GEN	125.77	151.63	143.17	129.10	*125.77*	no
KDIST and ...							
GHIER	KDIST + GHIER + CNT + SEM	115.18	151.63	136.42	122.93	*115.18*	no

Table 6.27. Pronominalisation in radio news: predictive power of semantic features, distance, and qualitative taxonomies of entity status. *italics*: last factor to be included, final AIC

HEAR, STAT4, STAT3, and SOURCE all yielded a higher AIC than KDIST in the first selection step, and once KDIST had been included in the model, the variables are not needed anymore. The only exception is DISC, corpus DLF-RE, which is included in the model PRO $\sim$ KDIST + CNT + SEM + GEN + DISC, and reduces the AIC from 125.77 to 125.70. Table 6.27 gives more details about the models with KDIST and GHIER. It shows that forward selection chooses roughly the same variables in the same sequence on both corpora. But on WBUR-LABNEWS, the resulting models account for less variation: the best model almost halves the AIC, while on DLF-RE, the best model reduces AIC by two thirds.

From these results, we can draw two conclusions: First entity status is the key to predicting pronouns; additional semantic information about the discourse entity does not greatly improve performance once we know its status. Second, in order to predict whether an entity will be pronominalised, we need to describe differences in the status of entities which have already been introduced into the discourse. This is what both the Givenness Hierarchy (GHIER) and the distance measure (KDIST) do, and it is the reason why they are so successful on the pronominalisation task. But what happens when we want to predict other aspects of form, such as the presence of a determiner, or whether speakers will use a definite description, or whether they will use an indefinite?

With regard to definites (Table 6.28), DLF-RE exhibits more variation than WBUR-LABNEWS: the AIC of the basic model (predict default) is 687.73 for the American, but 1093.1 for the German data. On WBUR-LABNEWS, neither semantic information nor entity status account for much of that variation: the largest reduction of the AIC we get is around 10%. On the DLF-RE corpus, however, we can manage to reduce the original AIC by roughly 20%, and the key to this is, again, the Givenness Hierarchy. As with pronominalisation, the semantics of the discourse entity is less important.

WBUR-LABNEWS							
	model		**AIC after including feature**				
	formula	AIC	KDIST	tax.	SEM	CNT	GEN
default	AIC = 687.73						
KDIST only	SEM + CNT + GEN + KDIST	635.07	*635.07*	n.a.	663.59	644.72	638.48
KDIST and ...							
GHIER	GHIER + CNT + SEM + GEN + KDIST	608.3	*608.3*	654.14	608.74	623.17	608.45

DLF-RE							
	model		**AIC after including feature**				
	formula	AIC	KDIST	tax.	CNT	SEM	GEN
default	AIC = 1093.1						
KDIST only	KDIST + CNT + SEM + GEN	1041.27	1061.8	n.a.	*1041.27*	1044.5	1041.5
KDIST and ...							
GHIER	GHIER + SEM + CNT + KDIST + GEN	850.24	850.79	916.19	857.45	865.68	*850.24*

Table 6.28. Definite descriptions in radio news: predictive power of semantic features, distance, and qualitative taxonomies of entity status. *italics*: last factor to be included, final AIC

For bare NPs (c.f. Table 6.29), this is different: here, the most important feature is almost always CNT, countability. The best taxonomy is, again, the Givenness Hierarchy, which yields the most parsimonious model on both corpora: BARE $\sim$ GHIER + CNT. Again, on DLF-RE, this model explains far more variation than on WBUR-LABNEWS: although both corpora start with the same AIC for the default model, the combination of GHIER + CNT manages to reduce the AIC by a third on DLF-RE, and by a fourth on WBUR-LABNEWS.

In order to determine which categories on the Givenness Hierarchy are mainly responsible for its performance, I inspected the coefficients of the fitted models. The most important categories are on one hand "active" (AC) and "in focus" (IF), on the other hand "referential" (REF) and "type identifiable" (TI). The first two categories are evidence against definites and bare NPs, while the second two tend to point to bare NPs, but rule out definite NPs.

Summary: The results I have presented in this section all point to one conclusion: when predicting the form of a referring expression, it is more important how accessible or salient a discourse entity is than where this accessibility comes from. This confirms the claims of Ariel (1990). Distance from last mention is a good measure of this accessibility, but it is severely limited in one respect: it does not categorise discourse-new entities further. But such categories are necessary, as well: That is the main message of the performance increases that the Givenness Hierarchy yielded across the board. In particular, it appears important to distinguish expressions which are referential or merely identify a type from those whose referent can be uniquely identified. To put it more drastically, the Givenness Hierarchy performs so well because it lumps identifiability and accessibility together. And these are the two main influences on the form of referring expressions that Chafe (1994) has identified.

But entity status does not tell the whole story. This is the message of the AIC values: A

WBUR-LABNEWS							
	model		**AIC after including feature**				
	formula	AIC	KDIST	tax.	SEM	CNT	GEN
default	AIC = 695.97						
KDIST only	CNT + KDIST + SEM + GEN	*551.76*	561.72	n.a.	558.36	603.79	*551.76*
KDIST and . . .							
GHIER	CNT + GHIER	530.37	no	*530.37*	no	603.79	no
SOURCE	CNT + KDIST + SOURCE + GEN	542.18	561.72	542.8	no	603.79	*542.18*
STAT4	CNT + KDIST + STAT4 + GEN + SEM	547.34	561.72	551.26	603.79	*547.34*	550.29
DISC	CNT + KDIST + SEM + GEN + DISC	*549.22*	561.72	*549.22*	558.36	603.79	551.76

DLF-RE							
	model		**AIC after including feature**				
	formula	AIC	KDIST	tax.	CNT	SEM	GEN
default	AIC = 681.23						
KDIST only	CNT + KDIST + GEN	551.14	555.70	n.a.	584.39	no	*551.14*
KDIST and . . .							
GHIER	GHIER + CNT	434.45	no	477.69	*434.45*	no	no
STAT4	CNT + STAT4 + GEN	543.18	no	548.93	584.39	no	*543.18*
STAT3	CNT + KDIST + STAT3 + GEN	543.48	555.70	549.56	584.39	no	*543.48*

Table 6.29. Bare NPs in radio news: predictive power of semantic features, distance, and qualitative taxonomies of entity status. *italics*: last factor to be included, final AIC

sizeable amount of the variation could not be explained by selected semantic information or by a judiciously chosen taxonomy of entity status. In particular, much of the variation in the distribution of definite descriptions is not accounted for. This shows how important it is to consider other factors than cognitive processing. For radio news, I would suggest two genre-specific pressures: the pressure to cram as much as possible into as few words as feasible, and a completely unjustified desire to avoid repetitions.

Finally, a word of warning about the validity of these results: I did not have access to training material for the annotation of the Givenness Hierarchy. Therefore my interpretation of the categories might not always agree with that of (Gundel et al. 1993), and when in doubt, I was certainly influenced by their previous results as presented in their paper. Second, many of the cells in the logistic regression models reported in this section were empty. This always causes numerical problems, especially if the amount of data is as small as it is here. These numerical problems affect the forward selection results, in particular when the model already contains two factors, and a third and fourth factor has to be added. I decided against explicitly specifying which cells could be expected to remain empty in the contingency tables. Since the analyses presented here are largely exploratory, I did not want my linguistic prejudices to bias the analysis too much.

6.4 Qualitative Analyses

In this section, I approach the texts from a completely different methodological angle: Instead of gathering statistics, I give detailed analyses of five texts. A set of four texts (Section 6.4.1) comes from DLF-RE. I chose the German texts because I am far more familiar with their content and context than with that of the WBUR-LABNEWS texts, which are heavily decontextualised. The analysis centres on the role of referring expressions in radio news, revisiting in detail some interesting patterns that have only been alluded to in the main text.

The fifth one (Section 6.4.3) is a short news report which has been analysed in detail by van Dijk (1985a) on the basis of his approach to text structure.

6.4.1 German Radio News

Two Dayton Stories

Throughout November 21, 1995, the war in Bosnia dominated the news. The audience of DLF were kept up to date on the latest developments in the Dayton peace conference. The whole day, conflicting statements and assessments poured in, all concerning the question: Will the talks be continued or not? The first story, given in Figure 6.4, explains the situation, while the second story, reproduced in Figure 6.5, reports on reactions of German politicians.

The first story is news because it concerns a turning point in a conflict that had been rich in personal drama (NEGATIVITY, PERSONALISATION) and that had captured the interest of both press and audience for four years already (CONTINUITY, RELEVANCE). This conflict was in Europe (PROXIMITY), and the crucial event is unfolding as the day goes by (RECENCY). Its topic is the US ultimatum to the Dayton delegates, but that topic is never expressed explicitly by a NP. Instead, it surfaces as the first sentence, the lead sentence. Sentence topics are almost impossible to identify; the news item appears as a succession of all-new sentences.

The second story is in a sense CO-OPTED by the first. It reports on how two prominent politicians, then-chancellor Kohl and the social democrats' international affairs expert Verheugen, see the current situation. Both are ELITE politicians, and can be supposed to represent the views of their party (ATTRIBUTION). The topic of that story is "reactions from German politicians to the Dayton situation", but that topic is never made explicit. Faithful news listeners can readily identify that type of item, which Zehrt (1996) calls *"Bonner" Meldung* (official news item), and which consists of reporting who said what on topic. When we analyse what each sentence is about, there are two alternatives. If we pursue the strand which consists of the comments on the Dayton peace talks, then the main contextual link of the first two sentences is the Dayton conference, the third introduces a new topic (a solution of the problems that only considers Bosnia), and the fourth sentence takes up that topic in the NP "ein derartiger Frieden" (*such a peace*; a peace based on such a solution of the problems). If we pursue the strand "who said what", then the first and third sentences introduce new topics (Kohl resp. Verheugen) and the second and fourth continue these topics.

In both texts, it is made very clear that they report statements by others. The second sentence in the first story is in past conjunctive, a sign of reported speech, and in the second story all main clauses contain either a full verb from the lexical field of saying or an equivalent idiomatic prepositional phrase ("nach seinen Worten", *in his words*). This leads to somewhat long and convoluted sentences.

After these more general stylistic comments, let us now turn to entity status. Both texts have strong inter-textual links: They follow each other immediately in the presentation, the second text comments on the facts reported in the first one. The NPs "der Bosnienkonferenz in Dayton" (*the Bosnia conference in Dayton*) and "der Verhandlungen in Dayton" (*the talks in Dayton*) in the first sentences of each text are both difficult to understand without the connection to other news texts that the audience have seen in the weeks and months before. The anaphoric definites in subsequent sentences are often synonyms, such as "der Delegationen aus Bosnien, Kroatien und Serbien" (*the delegations from Bosnia, Croatia, and Serbia*) . . . "die Vertreter der Konfliktparteien" (*the representatives of the warring parties*) in the first text, or "der Verhandlungen in Dayton" (*the talks in Dayton*) . . . "die Konferenz" (*the conference*) in the second text.

In the last sentence of the first text we even find an interesting pronominal anaphor. The "sie" (*they*) can both refer back to the delegates and to the American mediators. The second interpretation is reinforced by parallelism (two subject NPs), the first by recency (the object NP comes later in the sentence than the subject NP). The ambiguity is only resolved in the remainder of the sentence. It would not make sense to leave the American mediators alone in order to clarify open questions. On the other hand, the two possessive pronouns in the second text are relatively easy to resolve—both refer to the current main news actor.

What consequences do these observations have for entity status? The discourse entities are certainly not central to the structure of the news texts. In order to understand these texts listeners need knowledge about news schemata, in particular about types of news items. They also need background knowledge about current affairs; this helps them construct the initial descriptions of the discourse entities. Listeners also need that knowledge when accessing old entities, or they would not be able to decipher the synonyms or resolve ambiguous pronouns.

1. Die USA haben der Bosnienkonferenz in Dayton eine neue Frist bis zum Nachmittag gesetzt.[6]

2. Sollte bis sechzehn Uhr Mitteleuropäischer Zeit eine Einigung nicht erzielt sein, würden die Gespräche der Delegationen aus Bosnien, Kroatien und Serbien formell beendet.[7]

3. Weiter hiess es, die amerikanischen Vermittler hätten die Vertreter der Konfliktparteien am Verhandlungsort allein zurückgelassen, in der Hoffnung, dass sie untereinander die letzten strittigen Fragen klären könnten.[8]

Figure 6.4. War in Bosnia: the U.S. Ultimatum. November 21, 12:00, story 1

1. Bundeskanzler Kohl hat vor den Konsequenzen eines Scheiterns der Verhandlungen in Dayton gewarnt.[9]

2. In Singapur, der letzten Station seiner Asienreise, sagte der Kanzler vor dem Rückflug nach Deutschland, es wäre fatal, falls die Konferenz nicht zu einer positiven Regelung kommen sollte.[10]

3. Der außenpolitische Sprecher der SPD Bundestagsfraktion, Verheugen, vertrat heute früh im Deutschlandfunk die Ansicht, es reiche nicht aus, in Dayton eine Lösung der Probleme zu finden, die sich ausschließlich auf Bosnien beziehe.[11]

4. Ein derartiger Frieden wäre nach seinen Worten instabil.[12]

Figure 6.5. War in Bosnia: Comments of German politicians. November 21, 12:00, story 2

6.4.2 Two Rau Stories

In contrast to the two Dayton stories discussed in the last section, the stories in Fig. 6.6 and 6.7 are about a person, not about certain events. Both stories string together two apparently disparate news item:

- A certain Mr Horstmann has been named as a new secretary of state in North Rhine Westphalia (NRW),
- in the aftermath of a recent leadership crisis in the Social Democratic Party, Johannes Rau affirms that he will not leave politics.

In the first text the first story predominates. Although the second news item also reports on Horstmann's appointment, the emphasis is on Rau. His comments are set in a new context, that of the current leadership crisis in the Social Democratic Party. One of the main stories of the day is the vote of confidence that Rudolf Scharping, then leader of the Social Democrats in parliament, and now Secretary of State for Defence, has asked from his colleagues. What keeps both news items together is the main protagonist, Johannes Rau.

In the first text Rau is introduced as "Ministerpräsident" (prime minister) after the lead has stated the main news. In that lead, the function comes first, followed by the person who will fill it. The second sentence states the source, Rau, while the third and fourth sentences give background information: Who was the predecessor? Why did the post become free? When will the change take place? So far, the item has followed the classical news schema. But then, in the fifth sentence, the traditional schema is broken: the story reports Rau's answer to a question that must have been put at the press conference where he made Horstmann's nomination public, a question that challenges him to comment on the consequences that the current turmoil in the Social Democrats' leadership will have for him.

The lead of the second text focuses on Rau's reaction to that turmoil: He will dedicate himself to governing of his *Bundesland* now that his position in federal politics has weakened. This part of the news item again shows the classical structure: the lead (what Rau plans to do next), how that transpired (he said so before journalists), followed by one more sentence of Rau's opinions and a background sentence. The news about the new secretary of state is relegated to the final third of the story, and where the first news item connected Rau's reaction to the naming of Horstmann by a "Zugleich" (on the same occasion), in this item, the two parts are simply concatenated. The structure is also slightly permuted with respect to the classical schema. First comes the source, Rau, then the news, appointment of Horstmann, then the background (what will happen to Müntefering, why did the position become free?)

The referring expressions in both stories show patterns that are typical for radio news. News actors tend to be introduced together with their function. If they have several functions, journalists prefer that which is most pertinent to the current context. This explains why Rau is referred to as "Ministerpräsident" throughout the first story, which started as a report about an appointment he made in his function as prime minister of a *Bundesland*, whereas in the second story, he is introduced as vice president of the Social Democratic Party (SPD). In the fifth sentence of that story, when the focus switches to the appointment of Horstmann, Rau is referred to as "Düsseldorfer Regierungschef", *leader of the government in Düsseldorf*, where, as all adult Germans should know, the government of North Rhine Westphalia is located. The definite NP

is not only a welcome synonym, it also highlights the function of Johannes Rau that becomes relevant to the story now.

Of the six pronouns that occur in the two texts, two are possessives, and four personal pronouns. Both possessives have intra-sentential, subject antecedents. Three of the four pronouns occur in reported speech, in subject position, and refer to the speaker, Rau, who is also the main topic of the story. The last pronoun, which occurs at the start of the last sentence, is somewhat more difficult to resolve. It does not refer to Rau, the old topic, but to Horstmann. The discourse entity corresponding to Horstmann was introduced in the previous sentence, and it is highlighted by the marked word order (predicate before subject).

The anaphoric definites in these stories tend to compress much information into the space of a few words. For example, in both stories, all relevant information about Müntefering, Horstmann's predecessor, is packed into one NP followed by a relative clause. Other entities that are evoked frequently, be it directly or indirectly, are the Ministry of Labour, Health, and Social Issues (first text), and the Social Democratic Party (SPD). In both cases, I did not assign co-specification sequences to the mentions of these entities, because most of these mentions are indirect. In the second text, the SPD is mentioned twice as part of a compound (sentences 1 and 6), once, it has been elided (the "Mannheimer Parteitag" (*Mannheim party convention*), sentence 4, is obviously the party convention of the SPD), once it is referred to by the set of its members ("die Sozialdemokraten" (*the Social Democrats*), sentence 1), and only once as a party ("der SPD", sentence 3). The sequence of allusions to the Ministry is even more complex. Firstly, in sentence 1, it is evoked by an NP that refers to the function of its incumbent, then, in sentence three, it is evoked as part of a compound ("Ressort-" in "Ressortchef" *head of department*), and finally, in sentence 4, it is referred to by an uniquely identifying name (that is, uniquely identifying once you know that the Land which is evoked by the compound is the NRW). Listeners have to know about these mannerisms, else, it will be almost impossible to follow the news.

6.4.3 The Gemayel Text

In this section, I focus on a longer text that has already been analysed for macrostructural boundaries by (van Dijk 1985a). The aim of the analysis is to describe how discourse entities are maintained throughout longer stretches of text, and to determine how central the discourse entities are. The complete text is reproduced in Appendix A.1.

In my analysis of the Gemayel text, I will first address some problems with identifying referring expressions. Then, I concentrate on the form of subsequent mentions, and finally, I focus on the interplay between co-specification sequences and the superstructure of the discourse, as identified by van Dijk (1985a).

Identifying Referring Expressions: The text illustrates quite nicely four common problems with assigning referring expressions: times, coordinations, predicates, and idioms.

First, many dates have been labelled as referring expressions, although it is hard to imagine a continuation that refers back to these dates by a personal pronoun. In the Sortal Class labels defined in Appendix C, all of these NPs would be classified as Times, and Times as a class rarely get referred back to by NPs or pronouns. Regarding coordinations I followed the MUC guidelines (Hirschman and Chinchor 1997): If the coordinated NPs belong to co-specification

1. Neuer Minister für Arbeit, Gesundheit und Soziales in Nordrhein-Westfalen wird der SPD-Politiker Horstmann.[13]
2. Dies teilte Ministerpräsident Rau heute in Düsseldorf vor Journalisten mit.[14]
3. Damit tritt der einundvierzigjährige Horstmann die Nachfolge des bisherigen Ressortchefs Müntefering an, der das Amt des SPD-Bundesgeschäftsführers übernommen hat.[15]
4. Der offizielle Wechsel im Landesarbeitsministerium wird nach Angaben von Rau vermutlich in der nächsten Woche vollzogen.[16]
5. Zugleich trat der Ministerpräsident allen Spekulationen über seinen möglichen Rückzug auf Raten aus der Politik entgegen.[17]
6. Rau bekräftigte, er werde sich auf dem nächsten Parteitag wieder um das Amt des SPD-Landesvorsitzenden in Nordrhein-Westfalen bewerben.[18]

Figure 6.6. New Secretary of State in Northrhine Westphalia named. November 21, story 5, 13:30

1. Nach dem Führungswechsel bei den Sozialdemokraten will sich der stellvertretende SPD-Vorsitzende Rau auf sein Ministerpräsidentenamt in Nordrhein-Westfalen konzentrieren.[19]
2. Rau wies heute in Düsseldorf vor Journalisten alle Spekulationen zurück, er wolle sich allmählich aus der Politik zurückziehen.[20]
3. Im übrigen halte er seine Stellung in der SPD nicht für gefährdet.[21]
4. Auf dem Mannheimer Parteitag hatte Rau bei seiner Wiederwahl zum stellvertretenden Vorsitzenden lediglich achtzig Prozent der Stimmen erhalten.[22]
5. Der Düsseldorfer Regierungschef teilte mit, neuer Landesminister für Arbeit, Gesundheit und Soziales werde der einundvierzigjährige Politiker Horstmann.[23]
6. Er ist damit Nachfolger des bisherigen Ressortchefs Müntefering, der das Amt des SPD-Bundesgeschäftsführers übernommen hat.[24]

Figure 6.7. Rau will not leave politics. November 21, story 5, 14:30

sequences themselves they are marked as separate referring expressions, else, only the coordination itself is marked. The first case is illustrated in paragraph (14), where only the coordination is marked, the second in paragraph (18), where Begin and Draper are marked separately.

In contrast to the most recent version of the MUC guidelines I did not label predicating NPs as referring expressions. For example, in paragraph (5) the NP "Lebanon's president" refers to a function that Bashir Gemayel could have held, had he lived nine more days. To me the primary function of this NP is to predicate a potential function of Gemayel, not to refer to one of the constitutionally fixed positions in Lebanese society. The NP "a person elected president" in paragraph (6) is a borderline case: on one reading it can be said to be type identifiable, because it refers to the set of all people who have ever been elected president of Lebanon. On the other reading, it singles out Gemayel, because of all president-elects of Lebanon, he was the first to be assassinated. On that reading, the NP predicates a new property of Gemayel. Another tricky case are nominalised idioms. The collocation "in fear", paragraph (7), was not labelled as a referring expression, because it appears highly idiomatic to me. On the other hand, although "raised fears" in paragraph (5) is also a highly stereotypical collocate, I labelled "fears" as referring, because it is the object of "raise".

Tracking Discourse Entities: The co-specification sequences in the story show a number of interesting patterns. Table 6.30 protocols all sequences by paragraph number. The central entity is clearly Bashir Gemayel, the assassinated man himself, and his death and assassination are recurring themes. In the last paragraphs two other central entities surface: Menachem Begin and Yasser Arafat. Israel and Lebanon, two countries, also become more prominent in the last paragraphs, but in these sentences, they appear as agents and patients, not as locations.

Table 6.31 protocols the form of referring expressions for all discourse entities that were mentioned more than once. For each referring expression, I determined its form (pronoun, definite, indefinite, proper name, other), whether there were nominal or adjectival pre-modifiers, and whether prepositional phrases occurred as post-modifiers. A referring expression was labelled definite if either the definite determiner or a genitive occurred in determiner position.

News makers tend to be introduced with their full name and function; hence the five proper names with prenominal modifiers that occur as first mentions (c.f. Table 6.31). The definites concern events, such as Gemayel's death, his assassination, and the news of his death. Some discourse entities only surface sporadically in the text, such as the city of Jerusalem. It is only mentioned as the command base for high-ranking Israeli officials. In terms of a statistical model the presence (or rather: the first mention) of any Israeli official should predict a mention of Jerusalem, just as any mention of a statement by an U.S. president should increase the probability of the White House or Washington being mentioned. These probabilities encode journalistic conventions.

News actors are mostly referred back to by their names or by pronouns. Four of the ten anaphoric pronouns have their antecedent in the preceding sentence from the same paragraph, all others have intra-sentential antecedents. In the first third of the text, roughly up until paragraph (6), subsequent mentions are used to give new information about important news actors and locations. For example, we learn about Gemayel that he was the first president-elect of Lebanon to be assassinated, that he was a Maronite Christian, his age, and his political affiliation. About Lebanon, we learn that the country is "deeply divided". Such anaphoric definites that carry discourse-new or refresh relevant hearer-old information account for most of the

Class	Description	No. of paragraph
Person	Arafat	19, 20, 20
	Begin	16, 16(i), 16(i), 17, 18, 18(i)
	Draper	17, 18
	Gemayel	2, 4, 5, 5(i), 5(i), 6, 9, 10, 11, 11, 13, 13(i), 14, 14, 16, 17, 20
	Wazzan	9, 9(i)
Group	Israeli forces	1, 2, 4
	Israeli military command	2, 4
Event	G.'s assassination	2, 4, 9, 15, 16, 20
	G.'s death	5, 6, 9, 17
	news of G.'s death	9, 15
	the blast that killed G.	11, 12
Object	Gemayel's body	11, 11(i), 12
Location	west Beirut	2, 4
	Israel	8, 19, 19(i), 20, 20
	Jerusalem	2, 16, 17
	Lebanon	3, 5, 5(i), 8, 17, 18, 20, 20

Table 6.30. Co-specification sequences in the Gemayel text. For each entity, all paragraphs in which it occurs are protocolled. (i): intra-sentential antecedent

anaphoric definites in the text. The most frequent adjectival pre-modifier in the subsequent mentions is "criminal", and it is used for describing Bashir Gemayel's assassination.

Relation to News Schema: As van Dijk (1985a, page 85) remarks, the structure of this story is not at all linear. The first fifteen paragraphs are mainly strung together by the fact that they are all somehow related to Gemayel's death, while the last five focus on the reaction of two main players in the Middle East, Begin and Arafat. The first six paragraphs report on the latest events and refresh the most important background information. Paragraphs (7) and (8) are all-new sentences, contextual links are weak. They report consequences other than the invasion of Israeli troops. Paragraphs (9) and (10) are again only held together by the fact that both report reactions from high-ranking Lebanese politicians. We briefly regain referential continuity in paragraphs (13) and (14), which report on relevant aspects of Gemayel's history, but this is disrupted again by paragraph (15), which begins the section with verbal reactions from outside Lebanon. The three verbal reactions are contrasted by place: in the White House (Reagan) . . ., in Jerusalem (Begin) . . ., in Rome (Arafat).

All in all co-specification sequences do not contribute greatly to the coherence of the text. Its coherence is mainly guaranteed by the fact that readers are familiar with the news superstructure and the places that can be filled there.

	modifier	definite	pronoun	proper name	other
first mention	none	1	1	4	0
	prenominal	1	0	5	1
	post-nominal	1	0	0	0
subsequent mention	none	3	10	29	0
	prenominal	6	0	2	5
	post-nominal	6	0	0	0

Table 6.31. Forms of referring expressions in co-specification sequences

6.5 Summary

In this chapter, I analysed patterns of co-specification in corpora of (radio) news texts, focusing on DLF-RE, an excerpt from the Stuttgart Radio News Corpus (Rapp 1998), and WBUR-LABNEWS, an excerpt from the Boston University Radio News Corpus (Ostendorf et al. 1995). In contrast to earlier research on these corpora (e.g. Hirschberg 1993, Ross and Ostendorf 1996, Möhler 1998, Müller 1998), I explored to what extent results from media studies about the form and content of radio news can inform a linguistic analysis of such texts. Who gets referred to how and when in radio news appears to be not so much a linguistic but rather a sociolinguistic, even a political issue. The communication situation in the radio news domain is so complex that it is virtually impossible to classify the information in the news text into a dichotomy of given versus new information. Even when restricting givenness to the givenness of discourse entities, we face the problem that news "consumers", the typical addressees of radio news, are decidedly not homogeneous. For the purpose of this study, I mind-simulated a politically rather uninterested person called "John Doe" and annotated the texts according to how John would probably process them. I had to make informed (or, being true to my simulation, uninformed) guesses as to which news makers would seem familiar to him, and which ones he would be able to uniquely identify.

In the analysis, I focused on the connection between determiner type, pronominalisation, and entity status. Regarding the choice between pronoun and full NP, entity status is highly successful: Mesaures such as distance from last mention can explain more than 50% of the variation in the data. But when it comes to definite determiners or bare NPs, entity status is a lot less successful, covering only around 20% of the variation for bare NPs, and less than 10% for definites. I conclude from these observations that to make entity status the principle according to which languages structure their options for referring, as Ariel (1990) does, is at least questionable. This small study has clearly shown that other influences on the form of referring expressions are stronger. For radio news, this might be the constraint to cram much information into little time, which seduces editors to cram several propositions (in the sense of van Dijk 1980) into one single referring expression.

A more detailed analysis shows that successful entity status variables cover the two dimensions proposed by Chafe (1994), identifiability and activation. Too much detail appears to be harmful, as the investigation of the very detailed source-based scheme proposed in Section 5.2 has showed. Pronouns are very rare in corpora I looked at, and in contrast to the patterns found

in cross-genre corpora, such as BROWN-COSPEC, most of these pronouns have intra-sentential antecedents. Definites do not show any preference for anaphoricity or cataphoricity. Their distribution does not appear to be affected much by the linguistic context. Maybe in radio news, definite descriptions are the unmarked, default form of referring expressions. Indefinites, on the other hand, almost exclusively occur as first mentions. But in contrast to what is claimed in the literature (e.g. in the processing instructions of Givón (1992, 1995a)), most of the discourse entities that are introduced by indefinites are never mentioned again.

A detailed, qualitative analysis of four German radio news text and a brief journal article provided further insights into how referring expressions reflect entity status. In the German texts, referring expressions are carefully chosen not only to specify a discourse entity, but also to evoke the script that the addressee needs in order to process the information conveyed in the news item. In other words, the referring expressions not only specify discourse entities, they also set the scene for them. When the underlying story has been running for a while, the definite referring expressions tend to exploit intertextual relations between news items that have preceded them. The analysis of the newspaper article about the assassination of Bashir Gemayel shows that, at least in some genres, referring expressions are not as important for establishing textual coherence as linguistic theories would like to suggests. With this brief news report, its coherence comes mainly form the fact that it adheres to familiar patterns of news presentation; referring expressions are used to highlight relevant aspects of the persons that are talked about.

7 Pronominalisation

In the previous chapter, I have argued that if we want to examine linguistic correlates of entity status in a large-scale corpus study, we should measure entity status on the basis of co-specification sequences. In this chapter, I put the simple distance measure defined on page 111f. to the test in a detailed statistical cross-genre study of pronominalisation patterns in standard American English. The pronominalisation task can be defined by a simple question: In which contexts should we use a pronoun to mention a discourse entity? This task is an important subtask of the more complex task of generating referring expressions.

This chapter is structured as follows. In Section 7.1, I discuss influences on pronominalisation. Many of these influences are very difficult to measure, such as personal style, others are straightforward, such as agreement. Seven factors are identified that can be annotated reliably: agreement, sortal class, form of the antecedent, syntactic function, syntactic function of the antecedent, number of competing antecedents, and distance to last mention. The corpus we will use here, BROWN-COSPEC, is described in Appendix C. In Section 7.2, I examine whether these factors can be used to predict whether a referring expression should be pronominalised or not. In particular, we are interested in factors that perform well across genres. First, logistic regression is used to systematically test the predictive power of the factors (Section 7.2.1). A preliminary version of this section has been published as (Strube and Wolters 2000). We also examine whether the predictive power of the factors is robust with respect to genre (Section 7.2.2). Next, in Section 7.3 I examine how these factors fare when other approaches are used to learn the pronominalisation task. Two approaches are compared: automatic rule induction and exemplar-based learning. Finally, in Section 7.4 I discuss the results of this chapter in the context of previous research on generating referring expressions, and point out potential applications.

7.1 Influences on Pronominalisation

The influence diagram in Fig. 7.1 shows how linguistic and extralinguistic factors interact in the choice of linguistic forms.

In this figure, "Genre" stands for all constraints which come from conventions imposed by a discourse community for the discourse purpose of a text. Examples of genres are letters to the editor, academic research articles, or law texts. However, as soon as we apply this definition to standard representative corpora such as the Brown corpus or one of its mirror corpora such as the Lancaster-Oslo-Bergen corpus of British English (Johansson et al. 1986), we run into trouble, because the categories that were used for sampling the texts are a jolly mixture of genres, sub-genres, and domains (Lee submitted). A survey of the literature shows that there

are two alternatives: either re-classify the complete corpus, as (Kessler, Nunberg and Schütze 1997) have done, or make do with the existing categories. For our purposes we chose the second alternative, because it does not require us to design a scheme for determining the genre of arbitrary written English discourse, which is what Lee (submitted) ended up doing, and because the results are easy to replicate by others. The texts we chose are fully documented and described in Appendix C, where I also discuss the Genre variable in more detail.

"Content" stands for constraints imposed by the content. Domain-specific communication conventions are already covered under the heading of "Genre"; what I mean here is that there are preferred ways of talking about people, events, situations, states, etc. For example, humans tend to be mentioned as subject agents, while situations rarely, if ever, take the agent role, in particular if the writer is not in a mood for metaphor.

"Style" stands for the style of writing. Individual style leaves clear marks on a level as low as the frequency of function words (Mosteller and Wallace 1964, Holmes 1994). In their data, Henschel et al. (2000) found a constraint they call *repetition blocking:* Never use a pronoun two sentences in a row to refer to the same discourse entity. To me, this appears to be a genre-specific stylistic constraint. Both in newspaper copy and in the pedagogical descriptions of the MUSE corpus, it is important to get much information across efficiently, and new information about a discourse entity can be smuggled quite well into an anaphoric definite NP. This way, the writer avoids a full tensed clause. If we are to believe textbook writers such as Schneider and Raue (1998), computers should not necessarily be taught to mimic this questionable behaviour.

"Formal Constraints" are constraints imposed by grammar, while "Structural Constraints" are imposed by text structure, such as co-specification sequences, discourse segment boundaries, and relations between those discourse segments. Although the discourse entities themselves and their semantic properties belong to the domain of content, their status in the discourse is a structural constraint. Examples for the effects of these constraints have been given in Chapter 4.

Whether the reasons for using a pronoun instead of a full NP are the same for each genre, that still appears to be an open question. We know that there are large differences in the distribution of pronouns versus full NPs across genres. Table 7.6 demonstrates this for the BROWN-COSPEC-corpus, and Biber (1992) obtained similar results on his corpus of spoken and written British English. Fox (1987) took this observation one step further. She argued that the pronoun resolution strategies of a reader differ somewhat from those of a listener. Toole (1996) examined the distribution of referring expressions in four different genres, science fiction, book reviews, informal conversations, and current affairs interviews. She concludes that the distribution of referring expressions in all four genres follows the predictions of Ariel's (1990) Accessibility Theory, but this is difficult to verify, since her tables only relate to the complete corpus, never to single genres.

In this study, we examine whether we can "explain away" the genre differences if we feed our pronominalisation algorithms with the right features, features that can explain why one genre contains more pronouns than the other. For example, suppose we have a genre A that has many pronouns, of which most are first- and second-person pronouns, and a genre B with the same amount of third-person pronouns, but with no first- or second person ones. This difference can be explained by an Agreement feature that covers number. Whether we can find such a set of factors that predicts pronominalisation independent of genre, that will be a central question in the research reported in this chapter. The following Section 7.1.1 describes and

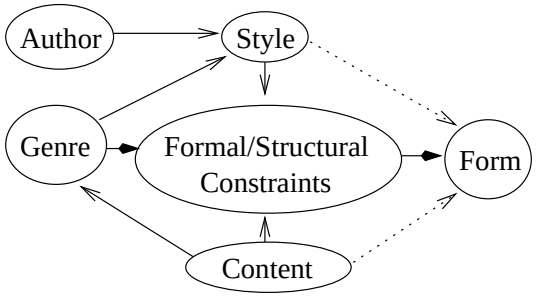


Figure 7.1. Constraints on the choice of linguistic forms. We are mainly interested in the dependencies signalled by arrows with filled heads.

motivates the factors that were used in our experiments, and Sections 7.1.2 and 7.1.3 present some preliminary quantitative data about the distribution of these factors and their connection to pronominalisation.

7.1.1 The Factors

We will model selected constraints by three types of factors here, which are summarised in Table 7.2:

formal **constraints** that come from the agreement value of the referring expression (AGREE) and its syntactic function in the sentence it occurs in (SYN). The possible values of these features are discussed in detail in section C.1.

content **constraints** that come from the semantics of the discourse entity. AGREE is already some sort of a content factor, because third person masculine and feminine pronouns are almost always used for persons. We supplemented this factor by sortal class (CLASS), which has been shown to be relevant for pronominalisation, and can be annotated reliably. In particular, we wanted to explore distinctions beyond the familiar [± animate] on the basis of a more detailed ontology. The sortal class annotations are discussed in more detail in Appendix C. I reproduce the annotation manual for the sortal class annotations in Appendix C.2.

structural **constraints** that come from the position of a discourse entity in a co-specification sequence. These factors are: distance to last mention (DIST), the number of competing antecedents (COMPANTE), parallelism (PAR), form of the antecedent (FORMANTE), and syntactic function of the antecedent (SYNANTE).

PAR is defined on the basis of syntactic function: a referring expression and its antecedent are parallel if they have the same syntactic function. COMPANTE is defined as the number of all discourse entities with the same agreement features that occur in the previous unit or in the same unit before the current referring expression. For DIST, we replaced the continuous, ordinal measure discussed in Section 5.3 by a categorical variable with four possible values, which are given in Table 7.1. This new variable allows us to cover both

-1	first mention
0	antecedent in same MCU
1	antecedent in previous MCU
2	antecedent more than one MCU ago

Table 7.1. Values of the variable DIST

NP level	
AGREE	Agreement in person, gender, and number
Values:	1 sg., 1 pl., 2 sg., 2 pl., 3 sg. masc., 3 sg. fem., 3 sg. neut., 3 pl.
SYN	Syntactic function
Values:	subject, object, PP adjunct, other
CLASS	Sortal Class
Values:	see Table C.2
Co-specification level	
SYNANTE	Syntactic function of antecedent
Values:	first mention, deadend, subject, object, PP adjunct, other
FORMANTE	Form of the antecedent
Values:	first mention, deadend, pronoun, possessive pronoun, demonstrative pronoun, definite NP, indefinite (with bare NP), proper name
DIST	Distance from last mention
Values:	no antecedent in discourse, antecedent in same MCU, antecedent in previous MCU, antecedent earlier
PAR	Parallelism
Values:	occurs with same syntactic function in previous sentence, yes / no
COMPANTE	ambiguity
Value:	number of competing discourse entities

Table 7.2. Overview of factors. All factors are categorical, COMPANTE is ordinal.

first mentions and subsequent mentions, whereas the ordinal measure was only well-defined for subsequent mentions. The Major Clause Unit (MCU) is defined in Definition 5.7, page 111.

The factors were selected on the basis of three criteria:

1. they can be derived from existing annotations
2. they can be annotated reliably
3. they can be annotated quickly - the more detailed the analysis, the slower will the annotators be

The second criterion was the reason why we did not annotate genericity, countability, or thematic roles. Poesio, Henschel, Hitzeman and Kibble (1999) have shown that it is very difficult to design annotation manuals which allow to annotate these features reliably. Thematic

roles are also eliminated by the third criterion. All linguists that give principled definitions of thematic roles (Dik 1989, Halliday 1994, Jackendoff 1992) base these definitions on either a classification of the main verb (Dik, Halliday), or on a decomposition of the main verb into semantic primitives (Jackendoff). A principled analysis of thematic roles would therefore require a deep and careful analysis of each proposition. It also introduces another confounding factor: not only can annotators disagree on thematic roles, they can also disagree on the verb classifications.

Although the way in which discourse is structured influences pronominalisation (see e.g. Fox 1987, Wiebe 1991, Wiebe 1994, Kibrik 1996), we will not investigate these influences here. There are several reasons for this. First before we can study how hierarchical structure affects pronominalisation, be it temporal, intentional, or attentional, we first need to investigate the linear base case thoroughly. Second, BROWN-COSPEC has not been annotated properly for discourse structure yet, and to devise an annotation scheme and implement suitable annotation tools would soon have led beyond the scope of this thesis. We could have simply estimated discourse segment boundaries by paragraph boundaries. For example, Zadrozny and Jensen (1991) treat paragraphs as building blocks of discourse, coherent units on the basis of which a formal discourse semantics can be specified. But the results presented in Table 7.10 demonstrate that the picture is not as straightforward as some of the literature seems to suggest. Since the relation of paragraphs to more conventional linguistic notions of discourse structure is in fact quite complex (Chafe 1994), I prefer to leave a study of paragraphs to a dedicated study of discourse structure, where paragraphs are compared with other approaches to discourse structure.

The aspects of discourse structure that are relevant for pronominalisation may vary from genre to genre, just as Fox (1987) assumed in her study. For some genres such as narrative, temporal relationships may be relevant, for others, such as police reports, formulaic building-blocks. Toole (1996) argues that for the purposes of analysis, it is crucial to use the same units for all genres. For her study, she chooses propositions and episodes as analysis units, following (Tomlin 1987a). Since her units are cognitively motivated, they do not depend as much on syntactic criteria as our MCUs. On the other hand, the properties of spoken and written varieties of a language can differ quite drastically, and for some data, such as conversations, we clearly need other methods of analysis than for e.g. legal documents. Fox adapted her analysis methods to the texts she worked on. This may make comparisons between genres more difficult, but her results are nonetheless valid.

Since we rely on co-specification sequences with identity links to determine the antecedent for a pronoun, there are several interesting types of pronouns we cannot cover: discourse-deictic pronouns, and plural pronouns that co-specify with two non-coordinated discourse entries. Fortunately, these pronouns are relatively rare in our data. We also do not distinguish between personal and demonstrative pronouns because demonstratives, again, are comparatively rare in our corpus. All three, demonstratives, discourse-deictic pronouns, and plural pronouns, are still hotly debated research topics. More and more researchers present corpus-based work on demonstratives (e.g. Botley 1996, Byron 1999) and discourse deixis (e.g. Eckert and Strube to appear, Eckert and Strube 1999). We expect that we can integrate these pronouns in future versions of BROWN-COSPEC.

The Brown corpus was chosen as the basis for our work because it is arguably the best studied corpus of American English there is. Numerous quantitative linguistic results have been published on it, and due to numerous revisions, its annotations are stable and reliable. We

Code	Description	
CF	Popular Lore	non-fiction, narrative or argumentative
CG	Belles Letters, Biographies	non-fiction, argumentative / expository
CK	General Fiction	fiction, narrative
CL	Mystery Fiction	fiction, narrative

Table 7.3. The genres in BROWN-COSPEC

Genre	# discourse entities	#seq.	% full NPs	% pronouns			
				total	first	second	third
CF	1223	125	80.4%	19.6%	1.4%	0.1%	18.1%
CG	1290	120	84.8%	16.2%	9.1%	0.2%	6.9%
CK	1071	113	63.8%	36.2%	8.1%	0.4%	27.7%
CL	954	170	64.4%	35.6%	5.3%	2.4%	27.8%

Table 7.4. Frequency of pronouns in genres. #seq: number of co-specification sequences. All percentages based on the total number of referring expressions.

did not include dialogue in this study because it is not clear how the units for the analysis of spoken language should be defined; even the MCUs (Definition 5.7, Section 5.3) that we will use here for written language represent a compromise between syntactic constraints, semantic structure, and ease of annotation, and other unit definitions clearly need to be investigated.

The genre definitions are taken directly from the Brown corpus categories. They are summarised again in Table 7.3 for convenience. Although, as we have seen, they have several disadvantages, the categories that we selected are relatively homogeneous, except for CF. Both CF and CG contain markedly fewer pronouns than CK and CL. The surprisingly high percentage of 16.2% for CG is mostly due to the first person pronouns in the expository texts (c.f. Table 7.4). In all other genres, most pronouns are in the third person.

In the following two subsections, I will report some preliminary quantitative analyses of the factors defined in the preceding section which should help us interpret the results in Sections 7.2 and 7.3. First, in Section 7.1.2, we will describe the distribution of referring expressions in BROWN-COSPEC. Then, we will examine the relationship between each of the values of these factors and PRO in more detail in Section 7.1.3.

7.1.2 Distances, Definites, and Pronouns

Table 7.6 shows that the distribution of referring expressions varies widely, even within genres. For example, texts CK25 and CK29 have the smallest number of sequences in the corpus, but this is due to the fact that each of these texts has two main actors which are mentioned in most sentences.

The texts from the two narrative genres, CK and CL, have fewer discourse entities than the others, but there are more and longer co-specification sequences and hence more referring expressions. Although the median sequence length does not vary greatly between texts and genres, the maximum sequence length does. A more detailed inspection of co-specification

Genre	Text	# entities	# sequences	% entities in sequences	sequence length median	sequence length maximum
CF		1226	128	10.44 %	3	73
	19	472	44	9.32%	2	11
	27	441	41	9.30%	2	5
	31	313	43	13.74%	4	73
CG		1290	120	9.30 %	2	67
	2	433	41	9.47%	2	16
	11	410	51	12.44%	2	32
	35	447	28	6.26%	3	67
CK		1081	123	11.38 %	2	123
	5	314	59	18.79%	2	96
	25	398	29	7.29%	2	109
	29	369	35	9.49%	2	127
CL		851	166	19.51 %	3	175
	4	292	51	17.47%	3	89
	6	247	49	19.84%	3	175
	22	312	66	21.15%	2.5	67
all		4448	537	12.07 %	3	175

Table 7.5. Distribution of Discourse Entities

Genre	Text	# ref. expr.	% definites	% indefinites	% pronouns	% 3rd person		
						sg. n.	sg. m./f.	pl.
CF		1725	22.50	47.83	19.59	2.96	7.65	1.68
	19	602	24.58	54.65	16.61	2.66	3.99	1.50
	27	506	24.51	55.14	13.83	5.93	1.38	2.57
	31	617	18.80	35.17	27.23	0.81	16.37	1.13
CG		1707	23.78	56.12	16.17	2.28	0.35	1.29
	2	544	31.99	57.17	7.35	3.31	0.18	0.37
	11	570	20.18	55.61	21.75	2.63	0.88	2.46
	35	593	19.73	55.65	18.89	1.01	0.00	1.01
CK		1848	17.05	38.91	36.15	4.33	13.58	1.68
	5	593	20.07	31.03	36.93	8.26	17.88	2.53
	25	624	15.87	43.91	35.90	3.21	18.75	1.60
	29	631	15.37	41.36	35.66	1.74	4.44	0.95
CL		1846	22.05	32.94	35.64	4.93	13.76	3.20
	4	587	25.04	31.18	35.26	3.92	17.72	3.92
	6	625	25.28	28.48	38.08	6.08	19.84	1.60
	22	634	16.09	38.96	33.60	4.73	4.10	4.10
all		7126	21.27	43.64	27.22	3.67	9.02	1.98

Table 7.6. Distribution of forms of referring expressions in BROWN-COSPEC. All percentages are based on the total number of referring expressions. The percentage of pronouns is based on both personal and possessive pronouns.

Genre	% first mentions as . . .								
	total			deadend only			tracking only		
	def.	in./bare	pro.	def.	in./bare	pro.	def.	in./bare	pro.
CF	25.94	64.27	2.60	25.77	65.57	2.10	27.34	53.13	6.25
CG	26.51	68.06	1.01	25.90	69.32	0.60	32.50	55.83	5.00
CK	23.68	63.27	4.26	23.28	65.45	3.97	26.83	46.34	6.50
CL	29.26	61.22	2.35	27.34	66.23	1.32	37.13	40.72	6.87
all	29.26	64.55	2.50	25.47	66.78	2.00	31.41	48.32	6.13

Table 7.7. Forms of first mentions in BROWN-COSPEC. Pronoun percentages refer to personal pronouns.

Genre	% occur as subsequent mentions		
	def.	in./bare	pro.
CF	18.04	4.48	86.15
CG	15.76	8.35	92.07
CK	18.73	4.87	90.36
CL	38.82	14.31	96.21
all	23.15	7.68	92.07

Table 7.8. Form of subsequent mentions in the BROWN-COSPEC-corpus

sequence lengths shows that narrative texts do not just have more sequences, they have more *long* sequences. This explains why we find more pronouns and less indefinites / bare NPs in CK and CLthan in CF and CG.

Definites: As Table 7.8 shows, most definites are first mentions. This confirms the results of e.g. Fraurud (1990). Indefinites and bare NPs appear to specialise in first mentions. They are especially frequent across genres for first mentions of discourse entities that are not accessed again, so-called deadend entities (c.f. Table 7.7). This might be due to a tendency that we have already observed in the Gemayel text, Section 6.4.3: Important entities that are new to the discourse are introduced by definite NPs which contain enough information to build a new, uniquely identifiable representation for that entity.

Now that we have some idea of the contexts in which non-anaphoric definites tend to be used, let us turn to anaphoric definites. According to Table 7.8, pronouns are the default anaphoric referring expressions. Tables 7.9 and 7.11 suggest that definite NPs are used instead of pronouns under two circumstances:

1. The antecedent occurs more than one unit ago. This holds for 68.67% of all anaphoric definites.

2. The next mention will occur in the next MCU. This holds for 48.47% of all anaphoric

Genre	Dist. 0				Dist. 1				Dist. > 1			
	total	% realised as . . .			total	% realised as . . .			total	% realised as . . .		
		def.	in./bare	pro		def.	in./bare	pro		def.	in./bare	pro
CF	11.94	6.80	0.97	88.83	9.68	14.37	9.58	59.28	7.30	25.40	15.08	15.87
CG	8.08	5.07	5.80	88.41	8.26	12.77	23.4	59.57	8.08	28.26	28.26	35.51
CK	16.45	0.99	2.30	94.74	15.53	2.79	3.14	89.20	9.52	27.27	10.80	41.48
CL	14.46	3.00	4.12	89.89	18.47	8.21	5.28	79.47	20.97	31.52	14.99	32.30
all	12.84	3.50	3.06	91.04	13.14	8.33	8.12	75.85	11.61	29.14	16.32	32.29

Table 7.9. Distance to last mention vs. form of referring expressions in BROWN-COSPEC. Percentage of mentions at distances 0,1, > 1 realised as definites, indefinites, pronouns.

definites. The tendency is even more marked for indefinites, again including bare NPs (67.85%).

The strength of these patterns differs from genre to genre. The texts from CF and CG show markedly more anaphoric definites than those from CK and CL. The median distance to last mention also varies greatly. It is highest for the two narrative genres, which also have more pronouns whose distance from their antecedent is greater than 1 MCU. This effect could be due to the first-person and third-person narrators in the narrative texts.

In the experimental literature, researchers have found that anaphoric definites cues the beginning of a new episode (Vonk et al. 1992). This prediction is not quite borne out by the corpus. For each paragraph in the texts, I determined whether there was a reference back to a discourse entity in the preceding sentence, and if so, whether it was realised by a pronoun or not. The results are summarised in Table 7.10. In fact, the only text where this hypothesis is confirmed is CF31. In all other texts there are either no cross-paragraph antecedents in the first sentences of paragraphs or these are realised by a pronoun. But if we examine these pronouns closer, we find that many refer to the main actor of a long discourse segment stretching over the complete or at least half the text. Alternatively, the pronoun can be a first-person narrator's "I". Finally, in text CG35, the speech, that pronoun is often a "we" which refers to "the American nation". In the text where we see the expected behaviour, CF31, the author discusses the behavior of several people in turn. On the basis of these results I would venture the hypothesis that anaphoric definites are more likely to be used episode-initially if the protagonists of the discourse segments change frequently.

Pronouns: Most of the subsequent mentions in the corpus are pronouns—across genres. As the distance to the last mention increases, the tendency to pronominalise decreases rapidly. The distribution of pronouns over distances shows a sharp fall, especially in comparison with the more gentle slope of definites (c.f. Figure 7.2). Although the distributions of definites, indefinites, and pronouns have the same mode, 1, their medians differ markedly (c.f. Table 7.11). Pronouns tend to have their antecedent in the same clause. But if they are themselves antecedent of a referring expression, that expression tends to occur in the next clause.

Pronouns predominate if the antecedent is in the same unit. This default weakens if the antecedent is in the previous unit, and it weakens even more for the non-narrative genres than

Genre	CF			CG			CK			CL		
Text	19	27	31	2	11	35	5	25	29	4	6	22
# paragraphs	15	12	26	13	10	39	46	5	14	45	29	13
cross-paragraph antecedents	4	0	**14**	1	3	21	12	4	13	21	14	5
pronominal	1	0	**2**	0	3	16	12	4	12	18	13	5

Table 7.10. Form of referring expressions with cross-paragraph antecedents. For each text, the table gives the number of initial sentences where referring expressions have cross-paragraph antecedents (second line) and the number of these referring expressions realised as pronouns

Genre	median distance to . . .					
	last mention			next mention		
	def.	indef.	pronoun	def.	indef.	pronoun
CF	1	2	0	1	1	1
CG	2	1	1	2	1	1
CK	7	2	1	1	1	1
CL	5	4	1	3	1	1

Table 7.11. Median distance to last and to next mention.

for the two narrative ones. The surprisingly large number of "long-distance" pronouns with a distance to the last mention of more than 2 units has several causes: First first- and second-person discourse entities are always referred to pronominally, no matter how long they have not been mentioned. Second, in all texts where long co-specification sequences occur, the associated entities are the main protagonist(s) of the story. Moreover, the story is told through the eyes of one of these protagonists. Such sequences account for most of the pronouns in the corpus, and for many long-distance pronouns, as well. The first-mention pronouns are mostly first- and second person pronouns, a small number has a clause-level antecedent.

7.1.3 Influence of Isolated Factors on Pronominalisation

After we have examined the distribution of referring expressions in our corpus, let us now reformulate the factors we will work with as random variables In this and the following sections, the factors SYNANTE, SYN, FORMANTE, PAR, CLASS, DIST and AGREE, whose values are discrete categories, will be modelled as a polytomous categorical variable, while COMPANTE is an ordinal variable. Pronominalisation is covered by the variable PRO. It can have two values: PRO, which stands for the event that a referring expression is realised as a pronoun, and NP, which stands for the event that a referring expression is realised as a full NP. What is the distribution of this variable, our target variable? Is it close enough to the binomial distribution to justify the use of logistic regression?

In order to examine the distribution of the random variable PRO, 1000 random samples of 100 referring expressions each were drawn from the corpus (with replacement), and in each sample, the number of pronouns was counted. Figure 7.3 shows the distribution of these counts. The mean number of pronouns in those random samples is $27.29 \approx 27$, the variance that was

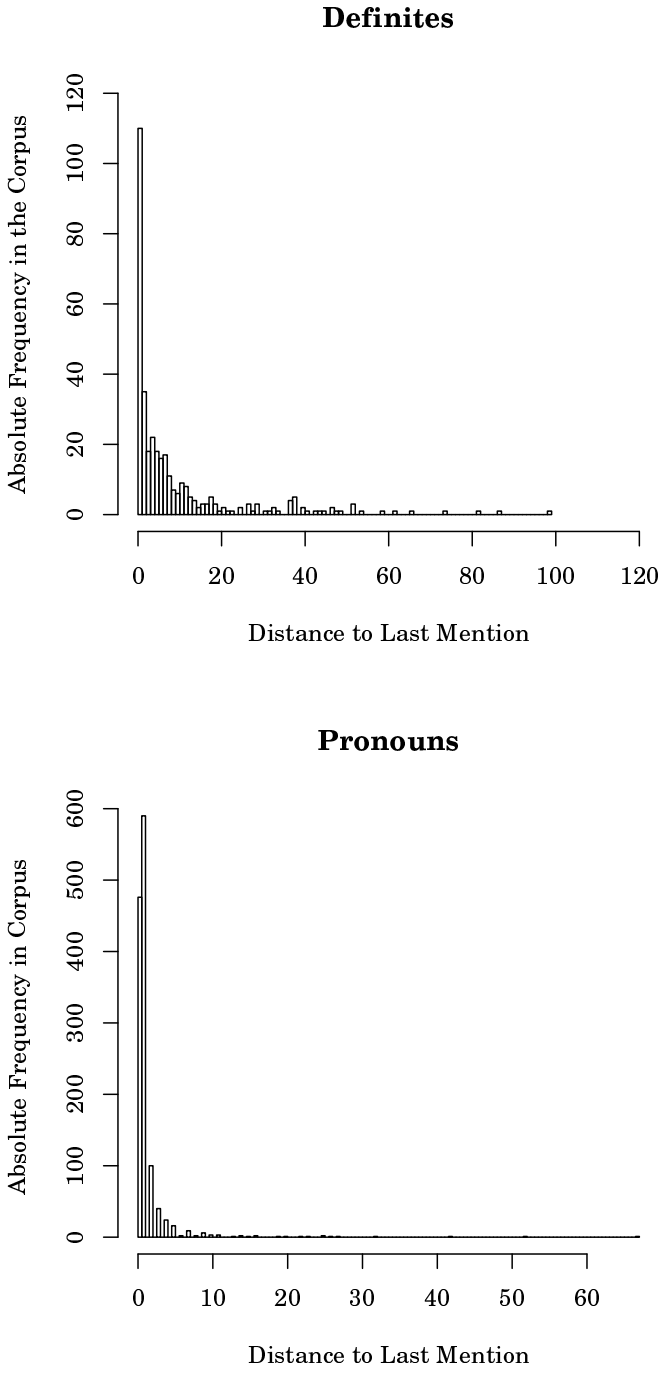


Figure 7.2. Distribution of distance to last mention for pronouns and definites in BROWN-COSPEC

	complete corpus	**Genres**			
		CF	CG	CK	CL
empirical mean $\bar{x}$	27.29	19.91	16.11	36.22	35.31
empirical variance s^2	21.00	16.55	13.75	23.70	23.33
estimated variance $\hat{\sigma}^2$	19.84	15.94	13.52	23.10	22.84
s^2 / $\hat{\sigma}^2$	1.06	1.04	1.02	1.03	1.02

Table 7.12. Mean, variance, and dispersion of PRO.

estimated from the data is 21.00. The corresponding means and variances for each of the genres is shown in Table 7.12.

The binomial distribution is the standard model for binary variables such as PRO. In terms of the binomial model (c.f. Appendix B.2), we count all pronouns as successes, and all full NPs as misses. p_{PRO} (or p, for short) is the probability that a referring expression will be realised as a pronoun. For a sample size of 100 referring expressions, the data yields an estimate of p=0.2691. The corresponding binomial distribution is plotted over the histogram in Fig. 7.3. The empirical distribution is somewhat broader than the theoretical one, and the counts around the mean are much more likely to occur than the binomial distribution would predict. The difference in the peaks of the two distributions may be due to the small number of samples that were used to estimate the empirical distribution. Many pronoun counts which are accommodated in the tails of that distribution are extremely unlikely to occur. The variance that was measured from the data also tends to be a little larger than the variance we would expect from a pronominally distributed variable. But the ratio of these two variances is always close to 1, as Table 7.12 shows — not enough to claim that PRO is seriously over- or under-dispersed. All in all, the binomial distribution fits PRO remarkably well, therefore we can safely use logistic regression, which assumes that the target variable, in our case PRO, has just such a binomial distribution.

Next, we examine the relationship between each of the factors and PRO. Table 7.9 demonstrates clearly that distance alone cannot account for all occurrences of pronouns in our corpus, although there are strong defaults: Almost all intra-sentential anaphora in the BROWN-COSPEC-corpus are pronouns, and most inter-sentential anaphora with the antecedent in the previous clause are pronominalised, as well. From a theoretical point of view, this is not surprising. First, there will always be contexts where both pronouns and full NPs are equally adequate, and where the choice between the two options is essentially stylistic. Second, as we have seen in Chapter 4, there are many factors apart from distance which influence how pronouns are interpreted—the syntactic structure of the sentence, the discourse structure of the text, the semantic structure of the propositions in which the referring expression is used, and so forth.

For the ordinal COMPANTE variable, the Kruskal-Wallis test was used, and the χ^2-test for the other, nominal, variables. We found statistically significant associations between PRO and each of the seven factors. These associations hold both for all referring expressions and for those that occur in sequences of co-specifying referring expressions. All of the tests were significant at the $p<0.001$-level, with the exception of PAR: for expressions that are part of co-specification sequences the effect of that factor is not significant.

For each factor, we determined whether some values are better cues to pronominalisation than others. The test we used is based on the fact that PRO has a binomial distribution. P(pro

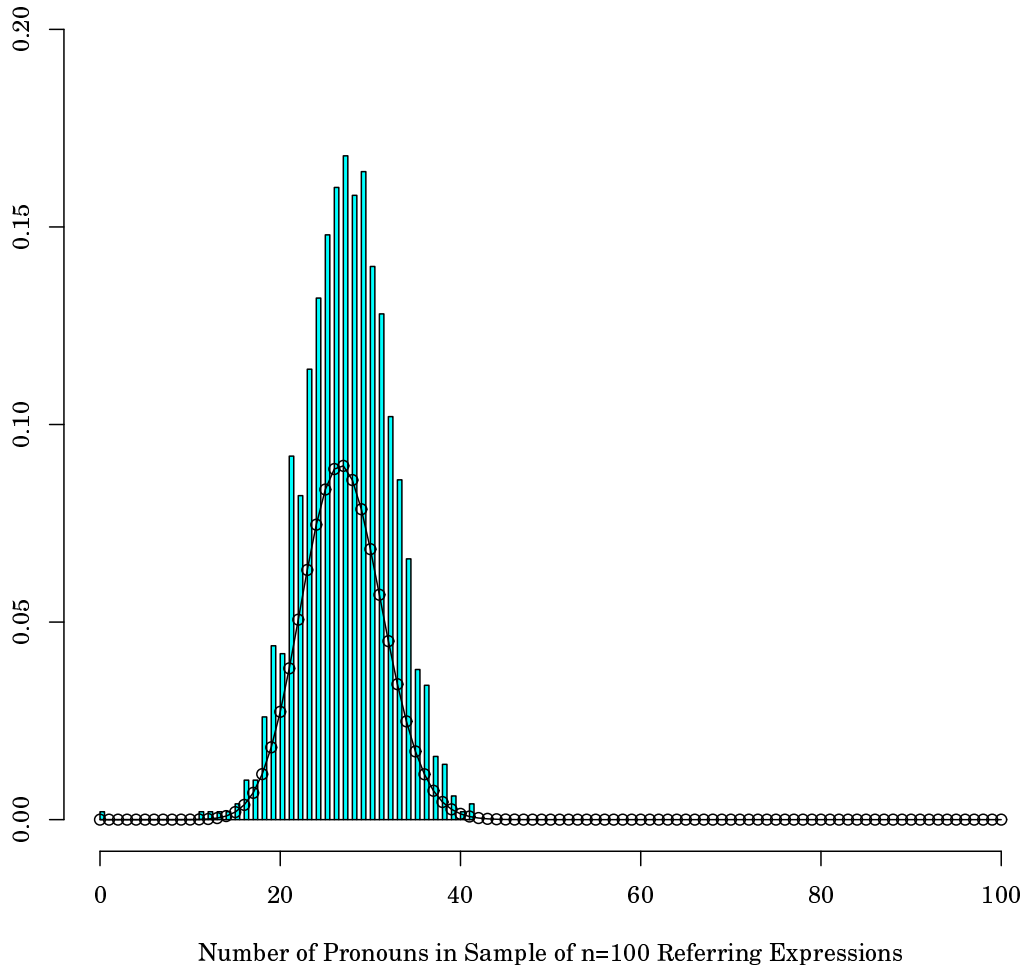


Figure 7.3. Distribution of PRO in BROWN-COSPEC. On the x-axis, we have the number of pronouns in a sample, on the y-axis, the relative frequency with which each of these numbers occurred in the set of 1000 samples. The histogram shows the distribution found in the data, the dots connected by a line the distribution which we would expect to find if PRO were binomially distributed.

$> x$;n,p) gives the probability that *x* or more referring expressions are pronominalised in a sample of size *n* if the "true" probability of finding a pronoun in the population from which the sample was drawn is p. In our case, n is the number of referring expression where factor X has value v, and p is the relative frequency of pronouns in the population from which we draw our samples. If P(pro $> x$;n,p) < 0.001, then a referring expression is significantly more likely to be realised as a pronoun if $X = v$. The rather strict significance level of 0.001 was chosen because we conduct many tests, and with a less strict significance level, chances are that some tests will give spurious results. We used three base data sets: the complete data set, all referring expressions in co-specification sequences, and all third person referring expressions in sequences.

The results can be summarised as follows:

AGREE: NPs referring to the first and second person are always pronominalised, Third person masculine or feminine NPs, which can refer to persons, are pronominalised more frequently than third person neuter and third person plural.

DIST: Pronouns are strongly preferred if the distance to the antecedent is 0 or 1 MCUs.

SYN,SYNANTE,PAR: Referring expressions are more likely to be pronominalised in subject position than as a PP adjunct, and referring expressions with adjuncts as antecedents are also pronominalised less often than those with antecedents in subject or object position. Pronouns are preferred as possessive determiners, and referring expressions that co-specify with an antecedent possessive pronoun are highly likely to be pronominalised. We also notice strong genre-independent effects of parallelism.

COMPANTE: COMPANTE has a significant effect as well: the median ambiguity for nouns is 3, the median ambiguity for pronouns 0. Closer inspection reveals that this is mainly due to first and second person and third person masculine and feminine pronouns.

The sortal classes show a number of interesting patterns (c.f. Table 7.14). Not only do the classes differ in the percentage of deadend entities, there are also marked differences in pronominalisability. There appear to be three groups of sortal classes:

1. *Person/Group* or [+animate], with the lowest rate of deadend entities and the highest percentage of pronouns. This is not only due to the first and second person personal pronouns.

2. *Location/Physical Object* or [-animate,-abstract], with roughly two thirds of all entities not in sequences and a significantly lower pronominalisation rate.

3. *Concept/Action/Event/Property/State/Concept* or [+abstract], with over 80% deadend entities. Within this group, *Action, Event,* and *Concept* are pronominalised more frequently than *State* and *Property*. *Time* is the least frequently pronominalised class. An important reason for the difference between *Location* and *Time*, which are both properties of situations, might be that *Times* are almost always referred back to by temporal adverbs, while locations, especially towns and countries, can be accessed via third person neuter personal pronouns as well as spatial adverbs.

SYN	all	CF	CG	CK	CL
subject	42.61	37.32	32.58	59.58	*40.40*
PP adjunct	4.86	2.84	2.19	13.30	12.56
object	14.40	9.80	6.37	*23.00*	*24.69*
other	*31.12*	32.36	35.19	*29.31*	*29.19*

SYNANTE	all	CF	CG	CK	CL
subject	72.60	72.31	70.27	87.93	66.10
PP adjunct	*32.86*	32.50	*23.08*	*45.45*	*38.46*
object	66.01	56.36	60.00	71.95	72.22
other	70.90	57.38	77.48	78.08	58.11

FORMANTE	all	CF	CG	CK	CL
definite NP	*31.89*	43.37	*24.36*	*46.97*	*23.64*
indefinite NP	50.12	61.05	35.34	66.25	*45.45*
personal pronoun	85.44	73.08	93.39	90.29	83.52
possessive pro.	83.48	67.95	89.65	86.31	85.22
proper name	47.51	46.51	*18.18*	58.62	*45.28*

Table 7.13. Pronoun frequencies for all values of syntactic function (SYN), syntactic function of the antecedent (SYNANTE), and form of the antecedent (FORMANTE). *italics*: not significant at $p<0.01$

Class	Person	Group	[+anim]	PhysObj	Loc	[-anim,-abs]
% deadend	17.28	46.09	20.86	65.46	63.25	34.90
% pronouns	63.39	28.41	59.04	10.17	5.65	9.42
% pron. in sequences	79.42	60.93	78.01	38.89	20.78	35.56

Class	Event	Action	State	Prop.	Concept	Time	[+abs]
% deadend	88.00	84.10	87.78	88.52	79.93	92.93	84.07
% pronouns	6.00	6.16	3.22	2.46	6.89	0.32	5.28
% pron. in sequences	60.00	45.28	39.13	42.86	42.01	7.14	41.30

Table 7.14. Pronominalisation of discourse entities from different sortal classes. % deadend and % pronouns are given relative to the total number of discourse entities in a class, the last row is relative to all non-first mentions of discourse entities from a class. Bold: no significant deviation from mean percentage over all sortal classes at $p<0.01$

Overall, the [± animate] distinction that the literature has been focusing on indeed appears to be the most important one. More complex ontologies, even if they are as small as ours, will not necessarily tell us more precisely when to pronominalise. If we examined anaphoric devices in general, we could determine why this is so: are locations and times just dispreferred antecedents for pronouns? Or are they less likely to be antecedents of anaphoric devices in general, no matter whether pronoun or adverb? To what extent is this pattern that some discourse entities from some sortal classes are much more likely to be referred back to than others language-specific? These answers are left to future work, for which other corpora may be necessary. Here, we focus on laying some groundwork concerning pronominalisation in American English.

7.2 Diagnostic Prediction: Logistic Regression

The factors defined in the previous section show strong statistical associations with pronominalisation—or, more precisely, with a binary variable PRO that codes for each referring expression whether it is realised as a pronoun or not. Now, we examine whether these factors can be used to *predict* whether a given referring expression will be pronominalised. In other words, we want to know whether the random variables that code these factors can be used to predict the value of an eighth random variable, PRO. In this section, we concentrate on a method which is particularly easy to analyse: logistic regression (c.f. Appendix B.3 and Andersen 1990, Agresti 1990, Lindsay 1995). Using logistic regression, we can find answers to the following three questions:

Question 1: How powerful are the factors we have defined? If a factor is powerful, it will account for a significant amount of the variation in the data set. Significance can be tested by various methods; we use both the F-test (McCullagh and Nelder 1983) and the likelihood ratio test (Andersen 1990, Agresti 1990) here, where the test statistic is equal to the deviance. The larger the amount of variation accounted for, the higher the F-score, the higher the reduction in deviance. The deviance measures the distance of the current logistic regression model to the saturated model, a model with perfect fit where each count is modelled by a separate parameter.

Such a model overfits the data and has no explanatory power at all. The power of a single factor F in isolation can be estimated on the basis of the model

(7.1) $$\text{PRO} \sim 1 + \text{F}$$

A model can be evaluated using the AIC, which stands for "An Information Criterion" (Akaike 1974). The lower the AIC, the better. The measure rewards a good fit to the data and punishes models with many parameters, which are likely to overfit the data. It consists of the deviance, which corresponds roughly to the amount of unexplained variation, and a term that incorporates the number of degrees of freedom of the model. The more degrees of freedom a model has, the more prone it is to overfit; the smaller the amount of unexplained variation, the better it fits the data.

If the model defined by Equation 7.1 has a low AIC, the factor F is powerful; it explains a great amount of the variation in the data. For example, DIST is clearly the most powerful criterion, while SYN and COMPANTE are relatively weak (Table 7.15).

When the model consists of several factors, the size of the contribution of a factor depends crucially on those factors that are already in the model: if two factors X, Y are not orthogonal, if they account for similar aspects of the variation in the data, then the effect of X will be much less dramatic when Y has already been included into the model, and vice versa. For example, FORMANTE covers much of the variation that DIST can explain (c.f. Table 7.17).

Question 2: Which factors are necessary for prediction? In order to evaluate the relevance of each factor, we use simple *forward selection* (Agresti 1990, Venables and Ripley 1997). We start with the most parsimonious model $\text{PRO} \sim 1$, which always predicts the default value, in this case, "full NP". Our aim is to find the model $\text{PRO} \sim 1 + \text{F}_1 + \ldots \text{F}_i$ with the lowest AIC. F_1 is the factor which reduces the AIC (Eq. B.9, page 268) of the base model $\text{PRO} \sim 1$ by the largest amount, F_2 is the factor that maximally reduces the AIC of $\text{PRO} \sim 1 + \text{F}_1$, and so on, until the AIC cannot be lowered anymore by adding a new term. If you compare Tables 7.15 and 7.16, you can instantly spot the factors that were added to the model first: they are printed in bold face, because they yielded the lowest deviance. AIC adds a penalty for the degrees of freedom to the deviance, but for the most powerful factors on each data set, that penalty was always smaller than the distance to the next largest deviance. The number of parameters of the models in Table 7.15 is df(F) (the degrees of freedom of the factor) + 1 (for the constant term); the resulting AIC penalty is thus $2 * (\text{df(F)} + 1) = 2 * \text{df(F)} + 2$.

Question 3: Is the influence of one factor on the target variable, PRO**, somehow mediated by another factor?** In this case, the interaction between the two factors will explain a significant amount of variation that the two factors in isolation cannot account for.

For example, in Table 7.18, we find that there is a large interaction between DIST and COMPANTE, the number of competing antecedents. The rule behind this is obvious: Do not pronominalise if there are competing antecedents in the same or in the previous MCU, even if distance to last mention is 0 or 1.

In particular, we would like to know whether we can build a genre-independent model of pronominalisation with the factors we have defined. In this context, two questions arise:

Data Set	AGREE	CLASS	COMPANTE	DIST	FORMANTE	PAR	SYN	SYNANTE
CF	1337	1372	*1616*	**805**	958	1472	1418	947
CG	821	1020	*1365*	683	**562**	1248	1197	676
CK	1471	1480	1968	**972**	1076	2060	*2200*	1122
CL	1579	1633	2044	**1212**	1226	1911	*2325*	1471
all	5363	5722	7186	**3802**	3953	6867	*7330*	4355
AIC	+16	+20	+4	+8	+18	+4	+8	+12

Table 7.15. Deviance of models PRO $\sim$ 1 + F. **bold:** smallest deviance for each data set, *italics:* largest deviance. Note: this table does not give the *reduction* in deviance, as most others do, but the deviance that remains unexplained by the model. The smaller the value, the better. The last line gives the term that you need to add to the deviance in order to get the AIC.

1. *Does genre influence the associations between a factor F and* PRO*?*
   If yes, then the interaction between genre and F in the following model should be significant:
   (7.2) PRO $\sim$ F + Genre + F:Genre
   Factors for which the interaction F:Genre is not significant are *robust*: their influence on pronominalisation, here modelled by the variable PRO, remains the same across genres.

2. *Can we predict* PRO *equally well for all genres?*
   If yes, then the percentage of correctly predicted pronouns should be around the same. A related question is: Is there any combination of features that yields optimal results for all genres? We will focus on these questions in section 7.3, when we discuss exemplar-based and rule-based approaches to pronominalisation.

7.2.1 Powerful Predictors

The most powerful factor, the factor that explains the largest amount of variation in the data, is clearly DIST. Table 7.15 shows that it is closely followed by FORMANTE. The two weakest factors, on the other hand, are SYN and COMPANTE.

The results of the forward selection experiments are summarised in Table 7.16. On the complete data set, the procedure yields the sequence DIST, AGREE, CLASS, FORMANTE, SYN, SYNANTE, COMPANTE, PAR. The sequences for the genres show interesting differences.

CLASS does not play a role in CG and CL, and FORMANTE, the second most powerful feature, is not needed for genre CF. We tested on the full data set whether it makes sense to replace CLASS by the factor NEWCLASS with the three categories [+animate], [+abstract], [-animate $\wedge$ -abstract]. The model PRO $\sim$ 1 + NEWCLASS has a deviance of 5918.6 and an AIC of 5924.6. The new factor performs somewhat worse than that for CLASS (Tab. 7.15). But when it comes to building a model, NEWCLASS is only inserted *after* FORMANTE and SYN. Although the additional amount of deviance it can explain is still significant ($p < 0.001$), the size of the effect has become small: 73.5. Table 7.14 suggests why: we have lost the distinction between Times and non-Times. Within the [+abstract] category, Times have special status; they

	complete corpus		CF		CG		CK		CL	
	Factor		Factor		Factor		Factor		Factor	
F_1	DIST	4542	DIST	902	FORMANTE	949	DIST	1446	DIST	1192
F_2	AGREE	800	AGREE	105	AGREE	128	CLASS	145	AGREE	262
F_3	CLASS	170	SYN	63	DIST	113	AGREE	60	FORMANTE	73
F_4	FORMANTE	110	CLASS	32	SYN	42	SYNANTE	*18*	COMPANTE	18
F_5	SYN	64	PAR	*3*	COMPANTE	*3*	SYN	*12*	n.a.	n.a.
F_6	COMPANTE	32	COMPANTE	*3*	n.a.	n.a.	COMPANTE	*6*	n.a.	n.a.
F_7	SYNANTE	*11*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table 7.16. Forward selection results. For each genre and for the complete data set, we give the sequence in which the factors were added as well as the reduction in deviance. The greater the reduction in deviance, the better. *italics:* improvement is not significant at $p<0.01$

are hardly ever pronominalised.

SYNANTE, which is also quite powerful, is only used twice, for the complete data set and for genre CK. In both cases, its contribution is not significant at $p<0.01$. This indicates that SYNANTE on the one hand and FORMANTE and DIST on the other cover much of the same ground. All three features code a distinction between first and subsequent mentions. SYNANTE and FORMANTE additionally distinguish first mentions of deadend entities, which are only mentioned once, from first mentions of discourse entities which are mentioned at least more than once. It is the additional information in the feature which counts here: the form of the antecedent as coded in FORMANTE appears to be more important than its syntactic function. PAR only occurs once (genre CF), and its contribution is not significant. Since it is also not very powerful, judging from Table 7.15, we will drop the feature in our future experiments. This yields the model in Eq. 7.3:

(7.3) CF: PRO $\sim$ DIST + AGREE + CLASS + FORMANTE + SYN + SYNANTE + COMPANTE

The model is purely additive; it does not include interactions between factors. This approach allows us to filter out factors which only mediate the influence of other factors, but do not exert any significant influence of their own. Results of a first evaluation of the full model are summarised in Table 7.20. The model can explain more than two thirds of the variation in the complete data set and predicts pronominalisation quite well on the data it was fitted on. The matter becomes more interesting when we examine the genre-specific results. Although overall prediction performance remains stable, the model is obviously suited better to some genres than to others. The best results are obtained on CG, the worst on CL (mystery fiction). In the CL texts, MCUs are short, a third of all referring expressions are pronouns, there is no first person singular narrator, and most paragraphs which mention persons are about the interaction between two persons.

In order to find out which values of the factors are particularly important for predicting the correct value of PRO, we examined the parameters of each value in the fitted model. All values of DIST have very strong weights in all models; this is clearly the most important factor. For AGREE, the first and second person are strong signs of pronominalisation, as well as, to a lesser degree, masculine and feminine third person singular. The most important distinction

excluded	fit		% explained variation						
	AIC	%correct	DIST	AGREE	CLASS	FORMA.	SYN	SYNA.	COMPA.
none	2686	92.6	54.4	21.1	5.7	3.8	2.3	*0.5*	1.1
CLASS	2785	**93.3**	54.4	21.1	n.a.	4.7	2.8	0.5	1.1
AGREE	2984	**92.6**	54.4	n.a.	**14.3**	6.2	2.7	0.6	1.1
DIST	3346	90.2	n.a.	35.8	6.1	**32**	3	0.8	*0.1*
DIST + CLASS	3443	90.2	n.a.	35.8	n.a.	**33.7**	3.4	0.8	*0.1*
DIST + AGREE	3597	89.6	n.a.	n.a.	**31.4**	**35.4**	3.1	0.8	*0.2*
AGREE + CLASS	3098	**92.6**	54.4	n.a.	n.a.	13.1	3.5	*0.5*	3.6
DIST + AGREE + CLASS	3739	89.4	n.a.	n.a.	n.a.	**52.6**	4.0	0.7	1.7

Table 7.17. Effect of leaving out any one of the three most important factors on model fit. The deviance value is the remaining deviance; it measures what is left unexplained by the model. The smaller, the better. *italics:* significance is $p < 0.05$, for all other factors, $p < 0.005$ or better.

provided by CLASS appears to be that between Persons, non-Persons, and Times. This also holds when the model is only trained on third person referring expressions. For singular referring expressions, personhood information is reflected in gender, and gender is coded in the agreement feature. But since English does not distinguish gender in plural forms, AGREE cannot replace CLASS for plural referring expressions. Another important influence is the form of the antecedent. The syntactic function of the referring expression and of its antecedent are less important, as is ambiguity.

In order to examine in more detail how important each factor is, we fitted the model from Eq. 7.3 on the complete data set, omitting one or more of the three central features DIST, AGREE, and CLASS. The results are summarised in Table 7.17. The most interesting finding is that even if we exclude all three factors, prediction accuracy only drops by 3.2%. This means that the remaining 4 factors also contain most of the relevant information, but that this information is coded more "efficiently", so to speak, in the first three.

How important is sortal class, which was, as the discussion in Appendix C shows, rather costly to annotate?

Well, remarkably enough, when sortal class is omitted, accuracy *increases* by 0.7%. The increase in AIC can be explained by a decrease in the amount of explained variation. A third result is that information about the *form of the antecedent* can substitute for distance information, if that information is missing. Both variables code the crucial distinctions between expressions that evoke entities and those that access evoked entities. Furthermore, a pronominal antecedent tends to occur at a distance of less than 2 MCUs. The contribution of syntactic function remains stable and significant, albeit comparatively unimportant.

Although DIST is clearly the dominant feature, there are considerable interactions between DIST and the six other factors. For each factor F, we construct a model PRO $\sim$ DIST + F + DIST:F and examine the reduction in deviance that each term yields. The results are summarised in Table 7.18. DIST interacts strongly with COMPANTE: the higher the ambiguity, the less likely it is that an entity will be pronominalised, regardless of distance to last mention. Interestingly, there is almost no interaction with the two syntactic factors, SYN and SYNANTE.

	Factor F		Interaction	
Name	df	Deviance	Deviance	Significance
AGREE	7	**800.2**	78.6	p<0.001
COMPANTE	1	327.0	**156.5**	p<0.001
CLASS	9	542.3	76.5	p<0.001
FORMANTE	7	498.5	60.4	p<0.001
SYN	3	205.8	20.7	no
SYNANTE	4	**147.8**	**8.5**	no

Table 7.18. Deviance of Terms F and DIST:F in model DIST + F + DIST:F. For all factors, the reduction of deviance is significant at p<0.001. The degrees of freedom (df) of DIST are 3, the degrees of freedom of the interactions are 3×df(F)

Name	# cat.	AIC	F-ratios			χ^2	df
			C_i	Genre	C_i:Genre		
DIST	4	3704.6	**1521.0**	30.3	4.3	492	9
FORMANTE	9	3889.7	560.0	16.6	3.8	674	24
SYNANTE	6	4263.5	815.0	31.8	3.1	928	15
AGREE	8	5260.5	377.0	32.2	3.1	1219	21
CLASS	10	5585.6	306.1	**44.0**	3.5	1860	27
COMPANTE	n.a.	7009.6	92.9	*2.2*	*2.9*	757	3
SYN	4	7164.5	344.9	25.0	**14.1**	2138	9

Table 7.19. Factors in pronominalisation. # cat: number of categories (except for COMPANTE, which is ordinal). For COMPANTE, χ^2(C_i,Genre) is Kruskal-Wallis χ^2. *italics:* value not significant at $p < 0.01$.

Both factors reduce deviance less than any of the other four. This suggests that most of the relevant information that they contribute is already implicit in the DIST variable.

7.2.2 The Influence of Genre

The frequency of pronouns in our data varies greatly with genre (Table 7.4). The distribution of our predictor features is also affected significantly by genre, as the χ^2-tests reported in Table 7.19 show. Tables 7.13 and 7.9 document the effect of three features on pronominalisation, both for the complete corpus and for each genre. DIST allows two very robust and general predictions: pronouns should not be used as first mentions, and anaphora within the same MCU should be realised as pronouns. Intersententially, preferences are more variable. The values of SYN, on the other hand, yield no clear predictions beyond a tendency to avoid pronouns in adjunct or direct object position; moreover, this tendency is subject to strong genre influences.

For each of the seven features, Table 7.19 shows the AICs of the models PRO $\sim$ C_i + G + C_i:G and the F-ratios of each model term. The most powerful feature is clearly DIST, with the

Model	**test data set**				
	CF	CG	CK	CL	all
MF	92.2	96.7	91.8	91.0	92.6 ± 0.0
MF without CLASS	**92.4**	**96.8**	**91.7**	**90.7**	93.0 ± 0.0
MP	91.9	96.1	91.4	90.1	92.6 ± 0.0
MPR	91.7	96.0	**91.7**	90.0	92.6 ± 0.0

Table 7.20. Performance of logistic regression models. **bold:** best model

lowest AIC and the largest F-ratio of all. COMPANTE, one of the weakest factors, is surprisingly robust: the interaction between that feature and genre is not significant. The additional amount of variation explained by Genre and COMPANTE:Genre is not even significant. Further logistic regression experiments show that once DIST has been included into a model, neither FORMANTE nor SYNANTE explains a large amount of the remaining variation. Instead, AGREE and CLASS become important terms. Although neither very robust nor very powerful, AGREE is the only feature that allows to predict the first and second person pronouns. CLASS is both less powerful than AGREE, and covers less genre-related variation (Genre F-ratio for AGREE: 32.2, Genre F-ratio for CLASS: 44.0). SYN is the least robust feature of all seven, with an F-ratio for SYN:GENRE of 14.1.

These results suggest two alternatives to the model MF in Eq. 7.3: A model which only takes into account the most powerful features (MP), and a model that combines powerful and robust features (MPR). The two models are defined by the following equations:

(7.4) MP: PRO $\sim$ DIST + FORMANTE + AGREE + COMPANTE

(7.5) MPR: PRO $\sim$ DIST + FORMANTE + SYNANTE + AGREE

Table 7.20 compares the performance of the full model MF, the full model without CLASS, MP, and MPR. The models were evaluated first by ten-fold cross-validation on the complete data set. For the cross-validation, the data set was divided into ten parts. In turn, each of these ten parts served as test set, while the other nine formed the training set. Table 7.20 reports mean and variance of the results on the test sets. We evaluated genre-independence by training on three genres and testing on the fourth. Table 7.20 shows that eliminating CLASS improves the performance of the full model. Apparently, the fine-grained class distinctions allow the model to overfit the training data. MP also generalises slightly better than MPR, although COMPANTE, which was the most robust predictor, is much less powerful than SYNANTE. The differences between the four models are all greater than two standard deviations.

7.3 Predicting Pronominalisation

We have seen that we need to supplement DIST by a number of other features if we want to model the pronominalisation patterns in the data adequately. Now, we will explore whether we can use our knowledge about influences on pronominalisation for improving the performance of classifiers which are to learn the classification task. We are particularly interested in finding

powerful and robust features: While *powerful* features should significantly increase average performance on the test set, *robust* features should consistently boost performance when training and test data is not from the same genre. We cannot expect that the best feature set for logistic regression will also be the best feature set for an arbitrary Machine Learning algorithm. The algorithms in the field differ too much in the properties of the data set that they can model.

In Machine Learning, there are two main approaches to feature selection: *Filtering* determines potential good factors on the basis of statistical data analysis, while in *wrapping*, the space of all possible feature combinations is searched for an optimal combination by training the classifier with many different feature sets. However, if we need to find an optimal feature set, we need to use wrapping (Kohavi and John 1998).[1]

In this section, we combine filtering and wrapping in order to determine feature sets which allow us to predict whether a referring expression should be pronominalised. We start with DIST, our measure of entity status, and the most powerful predictor according to the logistic regression models. COMPANTE is not included by default because its predictive power is rather low.

1. How well do the classifiers perform just on the basis of DIST? (This is the "filtering" component of feature selection: DIST was filtered from the original set of nine predictors on the basis of logistic regression analysis.)

2. Which combination of the six features SYN, SYNANTE, CLASS, FORMANTE, AGREE, and COMPANTE gives the greatest boost to this baseline performance? (This is the "wrapping" component of feature selection.)

3. How does the optimal combination vary across genres? Are there any features which are included particularly often? And what can the logistic regression models tell us about these features?

We experimented with two different approaches to Machine Learning: exemplar-based learning, represented by IB1(-IG) (Aha, Kibler and Albert 1991, Daelemans, van den Bosch and Weijters 1997) and rule induction, represented by RIPPER (Cohen 1995). The algorithms IB1(-IG) and RIPPER were chosen because they are widely used in the Computational Linguistics community. Both algorithms can deal with categorical as well as ordinal and interval-scaled features, although the similarity measure of IB1(-IG) is geared to categorical data. Therefore only the effect of categorical features can be reasonably compared across algorithms. The algorithms are described in more detail in Sections 7.3.1 and 7.3.2 together with the results.

We trained the algorithms using nine different setups for our data:

Setup A: 10-fold cross-validation on the complete data set. This setup shows how well the algorithms fare when they have no information about the genre of a text.

Setups CF**,** CG**,** CK**,** CL**:** 10-fold cross-validation on the 4 genre-specific data sets. These four setups reveal whether the algorithms perform better on some genres than on others. It also establishes optimal feature sets for each genre.

[1] It is beyond the scope of the thesis to discuss feature selection techniques in detail here.

Setups TCF, TCG, TCK, TCL: train on all data from three genres, test on all data from the fourth. These four setups test portability across genres.

We judged the performance of each algorithm according to four measures: classification accuracy, pronoun recall, NP recall, and the number of false positives. These measures are defined by the following set of equations:

$$\text{Accuracy} = \frac{\text{(\# correctly predicted pronouns and full NPs)}}{\text{(\# referring expressions in the corpus)}} \tag{7.6}$$

$$\text{Pronoun Recall} = \frac{\text{(\# correctly predicted pronouns)}}{\text{(\# pronouns in the corpus)}} \tag{7.7}$$

$$\text{NP Recall} = \frac{\text{(\# correctly predicted pronouns)}}{\text{(\# pronouns in the corpus)}} \tag{7.8}$$

$$\text{False Positives} = \frac{\text{(\# pronouns predicted instead of full NPs)}}{\text{(\# referring expressions in the corpus)}} \tag{7.9}$$

Since a pronoun contains less information about potential antecedents than a full NP, it is more difficult to resolve. Therefore it is better to have few false positives than a high pronoun recall.

We also explored how the parameter settings of the algorithms influence the performance of the resulting classifier. Ideally, one would perform such experiments using a validation set for testing which is different from both the training and the test data. Our corpus was not sufficiently large for this. As a result, the parameter settings are in a sense optimised for the test data we worked with.

7.3.1 Instance-Based Learning

The Algorithm: The basic algorithm is quite simple: Store every instance in the training data in an exemplar base. When new instances need to be classified, compare them to the instances in the exemplar base and assign the new instances the class of the most similar instance(s). Each instance consists of a description, which can consist of feature-value pairs, plus the class or category that has been assigned to that instance. For pronominalisation, we represent instances as feature-value pairs, and the classes are "Pro" and "full NP".

This simple algorithm lies at the heart of all instance-based learning (IBL) techniques, and many papers have proposed extensions and modifications to it. Aha et al. (1991) call that baseline algorithm IB1. All versions of IBL share the assumption that similar instances should belong to similar classes. They differ on three grounds (Aha et al. 1991, page 40):

- *the similarity function*: This function is crucial to the success of the algorithm. Many schemes have been proposed for weighting features according to their importance (Wettschereck, Aha and Mohri 1997). In IB1-IG (Daelemans et al. 1997), the features are weighted according to their informativity.

- *the classification function*: This function determines the class of the new instance on the basis of the similarity judgements from the similarity function. It can be extended in various ways: only take the most similar instance into account, organise a vote among the the k most similar instances, keep track of how often an instance from the exemplar base helped classify a new item correctly, and so on.

- *the instance base update function:* This function decides whether a new instance should be included into the exemplar base.

Exemplar-based approaches have been developed for many tasks such as grapheme-to-phoneme conversion (van den Bosch 1997), parsing (Bod and Scha 1997), word-sense disambiguation (Veenstra, van den Bosch, Daelemans, Buchholz and Zavřel 2000), and relative pronoun resolution (Cardie to appear). Daelemans, van den Bosch and Zavřel (1999) have shown that instance-based learning algorithms are particularly well suited for natural language learning, because in most such learning tasks, the categories do not form large clusters in the space of all possible instances. Instead, they are scattered into many clusters both small and large, corresponding to regularities, sub-regularities, and plain irregular instances. They show that a version of the basic IB1 algorithm where the contributions of each feature are weighted with its information gain (IG) outperforms IB1 on many natural language-related tasks.

Information gain is a quantitative measure of the information that a predictor variable P contains about the value of the target variable C. The *entropy* H(X) of a variable X codes the information it conveys. The more uncertain we are about the value of a variable, the more informative it is. The entropy is defined as follows:

$$H(X) = \sum_{v} -p(X = v) \log p(X = v) \tag{7.10}$$

H($C|P = v$) measures the uncertainty that remains about the value of the variable C if we know that variable P has taken on value v. To compute the information that P conveys about C, we just need to sum up the H($C|P = v$) for all values v of P. Now, we can define information gain as follows:

Definition 7.1 (Information Gain) *The Information Gain IG(P,C) describes the reduction in the entropy of H once we know the value of P.*

$$IG(P, C) = H(C) - \sum_{v} p(P = v)H(C|P = v) \tag{7.11}$$

The information gain of features with many values is potentially higher than that of features with few values, but many values are not necessarily beneficial to learning, since the more values a feature has, the higher dimensional the feature space, and the more difficult inductive learning becomes. Therefore we will use *gain ratio* (GR) instead of IG here. GR is given by the equation

$$GR(P, C) = \frac{IG(P, C}{H(P)} \tag{7.12}$$

Results: The instance-based learning results were gathered using the TiMBL package (Daelemans, Zavřel, van der Sloot and van den Bosch 1999). First, we explore how the parameter settings of the algorithm affect its performance. We concentrate on two adjustments: different neighbourhood sizes (parameter *k*, values 1,3,5) and gain ratio weighting as proposed by (Daelemans et al. 1997). In order to determine if these adjustments have any significant effects on our results measures, we conducted ANOVAs for all four measures and all nine setups. The influence of gain ratio weighting is always significant ($p<0.001$), as well as the interaction

	no Gain Ratio			**Gain Ratio**			**overall**				
	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5	no GR	GR
accuracy	**91.84**	91.47	91.34	91.86	91.94	**91.97**	**91.85**	91.71	91.66	91.55	**91.93**
false positives	**4.10**	4.61	4.81	4.05	3.95	**3.89**	**4.08**	4.28	4.35	4.51	**3.96**
pronoun recall	**84.92**	83.06	82.34	85.11	85.49	**85.69**	**85.01**	84.76	84.01	83.44	**85.43**
NP recall	94.43	94.62	**94.71**	**94.39**	94.36	94.32	94.41	94.49	**94.52**	**94.59**	94.36

Table 7.21. Performance of IB1 with/without Gain Ratio weighting and for different neighbourhood sizes

	CF	CG	CK	CL	TCF	TCG	TCK	TCL
accuracy - no GR	90.65	94.68	90.99	88.80	90.55	94.32	90.63	88.44
accuracy - GR	91.49	95.37	91.94	89.27	91.04	94.68	91.20	88.96
pronoun recall - no GR	69.32	76.64	85.69	84.98	78.37	79.83	81.88	85.39
pronoun recall - GR	75.53	83.11	88.58	86.68	82.68	86.69	83.69	87.23

Table 7.22. Average effect of gain ratio weighting for different genres

between presence of weighting and neighbourhood size ($p<0.001$) In most cases, we also find a significant effect of neighbourhood size. Table 7.21 illustrates the typical size and direction of these effects. Gain ratio weighting clearly improves the performance on pronouns: there are 0.5% less false positives and 2% more correctly recalled pronouns. The NP recall, on the other hand, drops slightly. The weighting also reduces performance variation when the neighbourhood size is varied. Indeed, without weighting, nearest neighbour consistently outperforms k=3 and k=5—except for NP recall. With weighting, this effect is exactly reversed: now, k=5 outperforms the two other sizes, except, again, for NP recall.

For the genre-specific data sets, we almost always find strong interactions between Genre and weighting ($p<0.01$). The only exception are the accuracy results on the tasks TCF, TCG, TCK, and TCL. Although weighting always increases performance, the size of these gains varies. Table 7.22 shows that the improvements are largest for genres CF and CG. For example, pronoun recall is increased by 6.86 percentage points for task CG, and by 6.47 for TCG. For CL and TCL, this gain dwindles to 1.7 and 1.84 percentage points, respectively. The effect of neighbourhood size is also mediated by Genre—sometimes, not k=5 gives the best results, but k=3 or k=1. For instance, the average accuracy on task TCG with weighting is 94.72% for k=1, 94.68% for k=3, and 94.63% for k=5. These results are interesting: Not only does each genre appear to require a specific combination of input features, but the optimal parameters of the learning algorithm also change with genre. In particular, some genres are more susceptible to small neighbourhoods than others.

In the following, we will discuss the results for IB1-IG with Gain Ratio weighting and neighbourhood sizes of k=5 and k=1. We will focus on the results for k=5, since that parameter setting results in a larger feature set. We only report accuracy, since the number of false positives and pronoun recall both correlate positively with accuracy.

Table 7.23 summarises the performance of IB1-IG on the nine tasks. For each factor, we determined whether it has a significant influence on accuracy using an ANOVA. Overall, the

exemplar-based approach performs as well as rule induction or logistic regression. It succeeds in capturing many of the relevant regularities that determine when a pronoun will be used instead of an NP. The most important factors are AGREE, FORMANTE, and COMPANTE. Including them improves performance significantly for six to seven of the nine tasks. Compared to results for k=1, given in Table 7.24, we find that k=5 only improves results for two genres, CG and CL, while the variation in performance increases. Increasing the neighbourhood size usually makes instance-based classifiers more robust and increases performance, as long as the neighbourhood is homogeneous enough (Duda and Hart 1973, Aha et al. 1991). As a consequence, more features can form the basis for the classification decision. The clear preference pattern of AGREE, COMPANTE, and FORMANTE that we find for k=5 is conspicuously absent for k=1, which does not tolerate additional features as well. Two features are used far more frequently with k=5 than with k=1: COMPANTE, the most robust predictor, and SYNANTE, the third most powerful factor. But SYNANTE almost never has a significant positive effect on performance, whereas COMPANTE is highly useful for 6 of the 9 tasks.

To get an idea of the type of mistakes that the instance-based classifiers make, we examined which instances in the test set are misclassified, and which feature values cause consistent, typical errors. Since we used 10-fold cross validation, the union of all 10 test sets gives the original data set. All percentages reported below were computed for the union of all test sets.

On task A, where all genres were pooled, the baseline values are acceptable: 4.6% of all full NPs are mistakenly classified as pronouns, and 14.5% of all pronouns are misclassified as full NPs. In general, there is a clear tendency to classify a given referring expression as a full NP. A more detailed analysis shows that the classifier is hindered by the strong distance-based defaults. 80.5% of all nouns whose antecedent appears in the same MCU are mistakenly classified as pronouns; for nouns with an antecedent in the previous MCU, that rate is 55.3%. Furthermore, only 22.3% of all first-mention pronouns are classified correctly. The data clearly does not contain enough information to offset these strong distance defaults. Another source of treacherous defaults is FORMANTE. In 55.5% of all cases where full NPs have a pronominal antecedent, the system generates a pronoun instead of a full NP. For AGREE, the effects of the defaults are not as strong. Because most third person neuter singular entities are full NPs, 41.7% of all third person neuter pronouns are not predicted correctly. We find the reverse for third person singular feminine: Here, most mentions are pronouns, and 26.4% of full NPs are mistakenly pronominalised.

The same patterns can be observed for the genre-specific tasks CF, CK, and CL, who have a high percentage of third-person pronouns. Most full NPs with antecedents in the same or previous MCU are confused with pronouns (CF: 74.7%, CK: 89.3%, CL: 73.2%). For CL, 79.3% of all full NPs with pronominal antecedents are misclassified, and for CK, 70.0%. The only exception is CG, the genre with the highest percentage of full NPs. Here, the misclassifications drop to 26.0% for full NPs with nearby antecedents, and to 35.3% for full NPs with pronominal antecedents.

7.3.2 Rule Induction

The Algorithm: Rule induction algorithms search for rules to perform the task they are supposed to learn. They extract the rules from the patterns they find in their training data. Some algorithms can also be bootstrapped with previous knowledge (Pazzani and Kibler 1992). A

IB1-IG—k=5									
Data Set	DIST only	DAFA	best overall accuracy						
			accuracy	features included					
				AGREE	COMPANTE	CLASS	FORMANTE	SYN	SYNANTE
all	90.1 ± 1.0	92.7 ± 0.8	92.7 ± 0.9	**X**	**X**	X	**X**	-	X
CF	91.5 ± 2.2	91.2 ± 1.0	92.9 ± 1.3	**X**	X	-	-	X	-
CG	91.6 ± 2.2	95.7 ± 1.5	96.6 ± 1.7	**X**	**X**	-	**X**	X	-
CK	90.7 ± 1.9	92.2 ± 1.9	93.0 ± 2.1	X	-	-	-	-	X
CL	86.8 ± 2.2	90.5 ± 1.9	90.9 ± 1.4	**X**	**X**	X	**X**	-	-
TCF	91.1	91.3	92.7	**X**	-	-	-	-	X
TCG	94.7	95.7	96.1	**X**	**X**	-	**X**	-	X
TCK	90.7	92.2	92.3	-	**X**	-	**X**	-	X
TCL	86.8	90.1	90.6	**X**	**X**	**X**	**X**	-	**X**

Table 7.23. Results for IB1-IG with k=5. For results that were obtained with 10-fold cross-validation, the standard deviation is given as well. DAFA: accuracy for feature set DIST, AGREE, FORMANTE, COMPANTE. **X** signals that the accuracy improvements achieved by including that feature are significant. An X indicates that the feature belongs to the feature set which yielded the reported maximal performance, a minus indicates that the feature was not included. Bold face: the influence of a feature on accuracy is significant (ANOVA, $p<0.01$)

IB1-IG—k=1									
Data Set	Dist only	DAFA	best overall accuracy						
			accuracy	features included					
				AGREE	COMPANTE	CLASS	FORMANTE	SYN	SYNANTE
all	90.1 ± 1.0	92.7 ± 0.8	92.7 ± 0.8	X	X	-	X	-	-
CF	91.4 ± 1.8	91.0 ± 1.0	93.0 ± 1.7	X	-	-	-	X	-
CG	95.2 ± 1.8	95.8 ± 1.4	96.2 ± 1.2	X	-	X	X	-	-
CK	90.7 ± 1.8	91.8 ± 1.7	93.0 ± 0.2	X	-	-	-	-	X
CL	86.8 ± 2.1	90.2 ± 1.8	90.6 ± 1.5	-	-	X	X	-	-
TCF	91.1	91.8	93.0	X	-	-	-	-	X
TCG	94.7	96.1	96.1	X	X	-	X	-	-
TCK	91.1	91.1	92.3	-	X	-	X	-	-
TCL	89.0	89.5	90.6	-	-	X	X	-	-

Table 7.24. Results for IB1-IG with k=1. For results that were obtained with 10-fold cross-validation, the standard deviation is given as well. DAFA: accuracy for feature set DIST, AGREE, FORMANTE, COMPANTE.

popular family of rule induction algorithms that includes CART (Breiman, Friedman, Olshen and Stone 1984) and C4.5/5.0 (Quinlan 1993, Quinlan 1996) represent these rules as decision trees. When the rules are represented as sets, the algorithm has to specify the sequence in which they have to be applied. The rules that are learned can be in various formats, procedural "if ... then" rules or Prolog-style Horn clauses, as in Inductive Logic Programming (ILP, Muggleton and De Raedt 1994, Bratko and Muggleton 1995, Wrobel 1996). Rule induction algorithms can be evaluated by their performance on unseen data, and by having experts examine the rules that they generated (Langley and Simon 1995). We will use both evaluation methods here.

The algorithm we use here, RIPPER (Cohen 1995, Cohen 1996), builds on Quinlan's (1990) FOIL, which learns first-order rules, and Fürnkranz and Widmer's (1994) IREP, an algorithm for generating small and concise rule sets. The algorithm generalises well: it performs as well as C4.5rules (Quinlan 1993) on a set of 37 benchmark problems, but is considerably faster.

RIPPER takes the most frequent category of the target variable as default, and tries to model the conditions under which the less frequent targets occur. If the target variable has more than two categories, it first induces rules for predicting the least frequent target category, then rules for the second least frequent, and so on, until it reaches the most frequent category, which becomes the default. In our case, the default is "full NP", the category that is predicted is "Pro". Full NPs are a natural default. Not only are they much more frequent in the data set than pronouns, they are also semantically richer and less ambiguous.

Further details of the algorithm are sketched in Figure 7.4. The rules that RIPPER generates have the format

(7.13) $$\text{TARGET} : -\text{CONDITION}_1 \wedge \ldots \wedge \text{CONDITION}_n$$

For categorical predictor variables X, conditions have the form X=x_i, where x_i is one of the possible values of X; for ordinal and interval-scaled variables, we have X$< \theta$ or X$> \theta$, where θ is a threshold from the data set. If only positive rules are allowed, then only these conditions can occur on the right hand side of a rule, if negative rules are allowed, as well, then these conditions can also occur in negated form, e.g. as X!=x_i. To classify a new instance according to the rules generated by RIPPER, the example is tested on each rule in turn, starting with the first one. The rules are ordered according to the number of examples they covered in the training data. If the conditions of a rule match the instance, it is assigned the target category. If no rule matches, the default category is assigned.

Since the form of RIPPER's rules is relatively simple, and since the optimisation pass does not check whether some of the generated rules are not in fact subsumed by a more general one, the rule sets that RIPPER generates will not be optimal for all learning tasks, and it can contain its share of redundancies, as we will see in the following paragraphs.

Results: The task of RIPPER was to induce rules for predicting the occurrence of pronouns. Rules were not forced to cover a minimal number of examples, and redundant rules were not allowed. We experimented with two parameters: the type of allowed rules (positive or negative), and the loss ratio, which is defined as the ratio of the cost of a false negative to the cost of a false positive. Since false positives (pronouns in the wrong place) are less desirable than false negatives (omission of pronouns), we only experimented with positive loss ratios. We did not perform an exhaustive search of the parameter space; instead, we trained RIPPER on the complete data set for the loss ratios 1 (default), 1.25, 1.5, and 2. Table 7.25 shows that permitting

Input: data set, number of iterations k
Processing:

while number of iterations < k **do**
generate rule set for each target category C (except default):
while (the rule set has not grown too complex)
and (there are still positive examples for category
C which are not covered by any rule in the rule set) **do**
split training data into growing and pruning data
generate a new rule:
while the rule still covers negative examples
add the condition with the highest information gain
on the growing data
prune the new rule:
delete any final sequence of conditions from the new rule
choose the version that maximises the pruning metric
on the pruning data set
remove all positive examples covered by the new rule from the training data
for each rule from the rule set
test whether it should be replaced by a new rule or
revised by adding more conditions

Output: set of rules ordered according to coverage

Figure 7.4. Outline of the RIPPER*k* algorithm. Positive examples are examples which have been classified under the target category C, negative examples are examples that have been classified under any of the other categories. The complexity of the rule set is measured in terms of total description length. The categories are ordered according to their frequency: the least frequent is modelled first, the most frequent becomes the default.

	only positive rules				positive and negative rules			
	l=1	*l=1.25*	*l=1.5*	*l=2*	*l=1*	*l=1.25*	*l=1.5*	*l=2*
median	91.96	**92.17**	91.83	91.70	**92.25**	92.00	91.97	91.56
minimum	**90.11**	**90.11**	89.55	89.73	89.90	**90.11**	89.80	88.55

Table 7.25. Accuracy for different rule types and loss ratios. Results are based on the mean performance in 10-fold CV runs over the complete data set.

RIPPER—negative and positive rules, test data									
Data Set	Dist only	DAFA	best overall accuracy						
			accuracy	features included					
				AGREE	COMPANTE	CLASS	FORMANTE	SYN	SYNANTE
all	90.1 ± 1.4	92.9 ± 1.2	93.1 ± 1.0	**X**	**X**	X	**X**	**X**	X
CF	91.5 ± 2.1	92.3 ± 1.4	92.7 ± 2.0	**X**	X	-	-	X	-
CG	91.6 ± 1.7	96.2 ± 0.9	96.7 ± 1.2	**X**	-	-	**X**	X	-
CK	90.7 ± 2.3	92.8 ± 1.6	93.1 ± 1.3	X	**X**	X	**X**	-	X
CL	86.8 ± 1.6	90.8 ± 0.8	91.0 ± 0.8	**X**	**X**	-	**X**	X	X
TCF	91.5	91.1	92.2	**X**	-	X	-	X	-
TCG	91.6	94.6	96.3	**X**	**X**	X	**X**	-	-
TCK	90.8	92.4	93.2	-	**X**	X	**X**	X	X
TCL	86.8	90.1	91.2	X	**X**	X	**X**	-	-

Table 7.26. Results for RIPPER, l=1.25, positive and negative rules, on the test set.
For results that were obtained with 10-fold cross-validation, the standard deviation is given as well. DAFA: accuracy for feature set DIST, AGREE, FORMANTE, COMPANTE. Bold face: the influence of a feature on accuracy is significant (Kruskal H-test, $p<0.01$)

negative rules tends to increase accuracy. The loss ratio is not as important for performance. Our small-scale search indicates that 1.25 is a good value.

The results for loss ratio 1.25 with negative and positive results are documented in Table 7.26. Including AGREE, COMPANTE, and FORMANTE almost always boosts accuracy. CLASS, SYN, and SYNANTE are also frequently included, but they are not as relevant for performance. Table 7.27 demonstrates quite nicely that the consistent pattern of AGREE, COMPANTE, and FORMANTE only emerges when we take generalisation into account—no matter if we test on data from the same genre(s) as the training data or if we test on different genres. If we merely take into account performance on the training data, the most important features are FORMANTE, SYNANTE, and AGREE. These are also the three most powerful features after DIST in Table 7.19. SYNANTE is only dethroned by COMPANTE when we determine the best features on the basis of the test data set.

The main increase in accuracy comes through the transition from DIST only to DAFA. For tasks A, CF, CK, and CL, the performance increase from DAFA to the best feature set is below 0.5%, for TCF, TCG, TCK, and TCL, it is somewhat higher - around 1%. It appears that for the T-tasks, some necessary rules are just not generated, because the relevant patterns are too weak in the training data. Table 7.28 shows that we find similar patterns if we only allow positive

RIPPER—negative and positive rules, training data										
Data Set	Dist only	DAFA	DAFS	best overall accuracy						
				accuracy	features included					
					AGREE	COMPA.	CLASS	FORMA.	SYN	SYNA.
all	90.1 ± 0.2	95.0 ± 0.7	96.1 ± 0.8	96.2 ± 0.6	X	-	X	X	X	X
CF	91.8 ± 0.2	94.8 ± 0.8	96.7 ± 0.4	96.7 ± 0.4	X	-	-	X	-	X
CG	91.8 ± 0.2	96.7 ± 0.4	96.8 ± 0.6	98.2 ± 0.4	X	-	-	X	X	X
CK	90.7 ± 0.3	94.5 ± 0.8	94.5 ± 0.7	95.3 ± 0.5	X	X	X	X	-	X
CL	86.8 ± 0.2	95.2 ± 0.6	95.4 ± 0.9	95.7 ± 0.1	X	-	X	X	-	X
TCF	89.7	95.4	94.7	97.2	X	-	-	X	X	-
TCG	89.6	95.5	95.4	96.6	X	-	X	X	-	X
TCK	89.9	95.9	96.1	97.1	X	X	-	X	X	-
TCL	91.3	95.6	96.6	96.8	X	-	X	X	-	-

Table 7.27. Results for RIPPER, l=1.25, positive and negative rules, on the training set. For results that were obtained with 10-fold cross-validation, the standard deviation is given as well. DAFA: accuracy for feature set DIST, AGREE, FORMANTE, COMPANTE. DAFS: accuracy for feature set DIST, AGREE, FORMANTE, SYNANTE.

RIPPER—positive rules, test data									
Data Set	Dist only	DAFA	best overall accuracy						
			accuracy	features included					
				AGREE	COMPANTE	CLASS	FORMANTE	SYN	SYNANTE
all	90.1 ± 0.5	92.6 ± 0.4	93.2 ± 0.3	X	X	X	X	-	-
CF	91.5 ± 0.7	92.3 ± 0.8	93.0 ± 0.5	X	X	-	-	-	-
CG	91.6 ± 0.6	95.8 ± 0.4	96.4 ± 0.3	X	-	X	X	X	-
CK	90.7 ± 0.8	93.0 ± 0.5	93.0 ± 0.5	-	X	X	-	X	-
CL	86.8 ± 0.5	90.8 ± 0.3	91.0 ± 0.2	X	X	X	X	-	X
TCF	91.5	91.1	92.1	X	-	-	-	-	X
TCG	91.6	94.4	95.8	-	X	-	X	X	X
TCK	90.7	92.6	93.4	X	X	X	X	X	X
TCL	86.8	90.4	91.3	X	X	X	X	-	-

Table 7.28. Results for RIPPER, l=1, positive rules, on the test set. For results that were obtained with 10-fold cross-validation, the standard deviation is given as well. DAFA: accuracy for feature set DIST, AGR, FORM, COMPANTE.

rules and set the loss ratio to 1. It appears that loss ratio and rule type are not as important as the preliminary experiments on the complete data set suggest. But if we compare the rule sets which are generated by the two versions, the story becomes more complex. Figures 7.5 and 7.6 document the rules that were generated most frequently by RIPPER. Only those rules are protocolled that occurred five or more times in all rule sets that were generated. A few complex rules have been merged into a single "or" rule, and when a rule head appeared in several rules with a few added conditions, these conditions are given in brackets. The most compelling difference between the two rule sets is maybe the first rule of Figure 7.5. It states something very much like Centering's Rule 1: The highest-ranked element of the list of forward-looking centres—here: the subject—will also be pronominalised in the following sentence, if no ambiguity will result and if that C_b was also a pronoun. This is a classical "continue" transition. Remarkably, that rule does not surface again in this clear form when negative conditions are allowed, as well. Both rule sets rely strongly on DIST. While the positive rules tend to couple DIST with either COMPANTE or FORMANTE, the other rule sets frequently use more complex conditions. Syntactic function is recurred to relatively rarely. Almost never do we find rules that consist only of conditions on SYN or SYNANTE.

The negative rules make much more use of these features than the positive ones. Typical conditions in which they occur are SYN/SYNANTE != "object" or SYN/SYNANTE != "PP adjunct". Essentially, these conditions express that the lower an entity is ranked in terms of its syntactic function, the less likely it is to be pronominalised. The ranking that is implicit here corresponds roughly to the grammatical function ordering of Centering (Grosz et al. 1995). The only value of CLASS that is used in the rules is *Person*. This suggests that the main relevant class difference for RIPPER is [± human], a feature which is very easy to label, once we have information about agreement values. For FORMANTE, we also find only two of the nine possible values in our rules: *pronoun* and *possessive pronoun*. This is explained by the finding documented by Table 7.11: once an entity has been pronominalised, it has switched to its "activated" state, and in that state, it is highly likely to be pronominalised again. Surprisingly, some very simple rules involving AGREE are almost never found, such as AGREE = "first person" or AGREE = "second person". Instead, such constraints are formulated indirectly as negative rules: if AGREE is not third person singular or plural, then pronominalise. Although some of the rules derived by RIPPER are surprisingly intuitive, and reflect linguistic theories quite nicely, the algorithm just does not catch some obvious generalisations that any first-year linguistics undergraduate could find. This result just serves to remind us of the Machine Learning truism that outcome of rule induction algorithms (and any machine-learning algorithm, for that matter) depends on the particular method for inducing the classifier, and on the representation of the input that the classifier gets.

Comparison and Evaluation: In order to evaluate the results of the three approaches, rule induction, exemplar-based learning, and statistical modelling, we compare their results with two baseline algorithms:

Algorithm A: Always choose the most frequent option (i.e. noun). This is a standard default rule.

Algorithm B: If the antecedent is in the same MCU, or if it is in the previous MCU and there is no ambiguity, choose a pronoun; else choose a noun.

```
PRO  :-  SYNANTE ="subject", COMPANTE <= 0, FORMANTE="pronoun"
PRO  :-  DIST = "same MCU"
PRO  :-  DIST = "previous MCU", COMPANTE <= 0
PRO  :-  DIST = "previous MCU", COMPANTE <= 2 or 3,
         AGREE= "third neuter"
PRO  :-  DIST = "previous MCU", FORMANTE = "pronoun"
         (, COMPANTE >= 6)
PRO  :-  DIST = "previous MCU", FORMANTE = "poss. pronoun"
PRO  :-  DIST = "previous MCU", SYNANTE = "subject",
         COMPANTE <= 4 or 6
PRO  :-  COMPANTE <= 1, DIST = "earlier than previous MCU",
         FORMANTE = "pronoun" or "possessive pronoun"
PRO  :-  COMPANTE <= 1, AGREE = "1st pl."
PRO  :-  COMPANTE <= 1, AGREE = "2nd sg."
PRO  :-  FORMANTE = "pronoun", COMPANTE <= 0
PRO  :-  FORMANTE = "possessive pronoun"
         (,COMPANTE <= 0 or DIST = "previous MCU" or AGREE = "1st pl.")
```

Figure 7.5. Frequently used rules for RIPPER, full data set, best feature set DIST, FORMANTE, SYNANTE, AGREE, COMPANTE, positive rules only. Only those rules are protocolled that are generated more than 5 times. The rules specify conditions for pronominalisation. "Full NP" is the default class.

```
PRO  :-  DIST = "same or previous clause" (, COMPANTE<=0)
PRO  :-  DIST != "first mention" (, AGREE!= "third neuter or plural" or
         AGREE = "first plural")
PRO  :-  DIST != "first mention", COMPANTE<=0,
         FORMANTE= "pronoun"
PRO  :-  DIST = "previous clause", SYN != "PP adjunct",
         FORMANTE="pronoun"
PRO  :-  CLASS = "Person", AGREE != "third person" or
         AGREE = "second singular"
PRO  :-  CLASS = "Person", FORMANTE= "pronoun",
         (, COMPANTE <= 0 and/or SYNANTE != "object")
PRO  :-  FORMANTE != "deadend", AGREE= "first sg./pl."
PRO  :-  FORMANTE = "pronoun", COMPANTE<=0
```

Figure 7.6. Frequently used rules for RIPPER, both positive and negative rules, best feature set. Only those rules are protocolled that are generated more than 5 times. The rules specify conditions for pronominalisation. "Full NP" is the default class.

Approach	test data set				
	CF	CG	CK	CL	all
Algorithm A	80.4	83.8	63.8	65.4	72.8
Algorithm B	91.1	93.0	88.6	84.7	89.4
Model	92.2	96.7	91.8	91.0	92.6 ± 0.0
Model without CLASS	92.4	**96.8**	91.7	90.7	93.0 ± 0.0
IB1-IG, k=1	**93.0**	96.1	92.3	90.6	92.7 ± 0.7
IB1-IG, k=5	92.7	96.1	92.3	90.6	92.7 ± 0.9
RIPPER, positive rules	92.1	95.8	**93.4**	**91.3**	$\mathbf{93.2 \pm 0.3}$
RIPPER, negative and positive rules	92.4	96.3	93.2	91.2	93.2 ± 0.3

Table 7.29. Results of algorithms vs. models on test data in % accuracy.

Algorithm B is based on the most powerful predictor, DIST, and the most robust predictor, COMPANTE. It also takes into account the strong interaction between the two predictors that is also evident from the RIPPER rules.

Table 7.29 summarises the results of the comparison. To determine the overall predictive power of the model, we used 10-fold cross-validation. Algorithm A always fares worst, while algorithm B, which is based mainly on distance, the strongest factor in the model, performs quite well. Its overall performance is 3.2% below that of the full model, and 3.6% below that of the full model without sortal class information. Nevertheless, for all genres, the statistical models, IB1-IG, and RIPPER outperform the simple heuristics. Excluding sortal class information can boost prediction performance on unseen data by as much as 0.4% for the complete corpus. The apparent contradiction between this finding and the results reported in the previous section can be explained if we consider that not only were some sortal classes comparatively rare in the data (Property, Event), but that our sortal class definition may still be too fine-grained. For the two narrative genres, CK and CL, which contain far more pronouns than CF and CG, the improvement over the baseline (Algorithm A) is largest: between 25 and 30 %. Performance on CF and CG increases by about the same amount: we have 12.6% for CF (best algorithm IB1-IG, K=1) versus 13% for CG (logistic regression). The performance of RIPPER and IB1-IG for different parameter settings is basically stable. RIPPER enjoys a slight advantage on the two narrative genres, which have more pronouns. Apparently, RIPPER can model some of the more intricate patterns of pronominalisation in these texts better than IB1-IG, but because the vast majority of cases are covered by straightforward rules, which rely on strong defaults for certain feature values, both algorithms are almost equivalent.

The results on the datasets TCF, TCG, TCK, and TCL suggest that it is not possible to find a single feature set that performs equally well on all genres. The inconsistent performance of DAFA, a feature set that combines the three best features so far, AGREE, FORMANTE, and COMPANTE, with DIST, corroborates this finding. This result is not surprising: For finding an optimal feature set, classifiers have to be trained with different combinations of features; "offline" pre-selection tends to give worse results (Kohavi and John 1998). Therefore we cannot present an off-the-shelf algorithm for pronominalisation. Instead, we propose a comparatively fast off-the-shelf *strategy*: Annotate a representative set of texts with co-specification

sequences, determine the form of the referring expressions, determine agreement, if the texts contain first- and second person pronouns, and train a classifier on that data using distance information plus all possible combinations of COMPANTE, AGREE, and FORMANTE to find the best feature set for your classifier. Features that are comparatively expensive to annotate, such as the two syntactic features SYN and SYNANTE, and especially CLASS should be left out, if the main goal is to bootstrap the generation algorithm from a corpus. Later on, these features can be incorporated into the full algorithm when they become available from the generation module.

7.4 Discussion

In this section, I put the results of this chapter in the wider context of related research and potential applications. First I briefly discuss related work on pronoun generation and machine learning (Section 7.4.1). Then, I point out some applications (Section 7.4.2)

7.4.1 Related Work

Centering (Grosz et al. 1995) already provides a rule which decides for some referring expressions whether they should to be realised as a pronoun, namely Rule 1 (Definition 4.3, page 79: The highest ranking forward-looking centre of utterance U_{n-1} that is realised in utterance U_n can be realised as a pronoun (for further explorations of Rule 1, c.f. e.g. Kibble 1999, Kibble and Powers 1999).

Another strand of research on generating referring expressions is based on the Gricean Maxim of Quantity, which exhorts communicators to make their contributions as informative as necessary. This maxim implies that referring expressions should contain as much information as the addressee needs to identify the specified discourse entity. The Incremental algorithm of Dale and Reiter (1995) assumes that the entities in the domain are described by sets of properties. These properties are characterised by attribute-value pairs. Attributes are ordered on a preference hierarchy depending on which characterisations people are more likely to use. For example, absolute properties such as colours are preferred over relative attributes such as sizes. For some attributes, a subsumption hierarchy is defined on their values. Each entity has at least one "type" attribute. When a referring expression for an entity has to be generated, the aim is to generate a parsimonious definite description that rules out all detractors and makes the intended entity uniquely identifiable. The incremental algorithm passes through each of the attributes according to the preference ordering. For each new attribute, it determines the best value, that is the value which does not rule out less distractors than the values it subsumes. If the description that has been generated so far plus the new attribute-value pair uniquely identify the entity for which a referring expression is to be generated, the algorithm stops. Krahmer and Theune (1999) extend the Incremental algorithm by an explicit notion of *salience*. They argue that a definite description is sufficiently precise iff there is exactly one most salient object that corresponds to that description.

In the algorithm of Appelt (1985), referring expressions are generated so that they fulfil certain communicative goals. His application domain are task instruction monologues. When a new goal from the task plan is to be integrated into the monologue, a set of critics test how it can

best be connected to what has already been generated. For example, the addressee needs to use a wrench. Next, he needs to be informed that the wrench is in the toolbox. Instead of generating two sentences "Use the wrench. The wrench is in the toolbox.", the algorithm integrates them, instigated by one of the critics, into the sentence "Use the wrench in the toolbox.". Jordan (2000) explores systematically on a large corpus of spoken task-oriented dialogs how communicative goals influence the form of referring expressions. She tests several rule-based algorithms on her corpus, but does not experiment with machine learning methods.

The Machine Learning approach to pronominalisation that we have discussed in Section 7.3 is not geared to any of these high-level goals. It incorporates a very rough notion of salience: the shorter the distance to last mention, the more salient a discourse entity will be. We have also coded the syntactic function of the antecedent; syntactic function is the classical basis for computing the order of the forward-looking centre list in Centering (c.f. Section 4.3.2).

McCoy and Strube (1999) pursue yet another route. They want to know which information that a generation system needs in order to decide between a pronoun and a full NP. They explore distance from last mention in sentences, temporal discourse structure, and ambiguity. A pronominal reference is defined as *unambiguous* if it can be resolved successfully by the algorithm of Strube (1998). Only those pronouns are generated which can be resolved unambiguously. Pronouns are blocked if a change in temporal structure occurs between the anaphor and its antecedent. The algorithm was evaluated on a corpus of three reportages from the New York Times and achieved an accuracy of 84%. The solution of Henschel et al. (2000) is modelled along the lines of (McCoy and Strube 1999), with two important differences: they use predetermined discourse segment information instead of temporal structure, and they introduce a stylistic "repetition blocking" that prohibits chains of pronominal references. They get slightly better results than McCoy and Strube (1999) on the New York Times corpus.

The only dedicated machine learning approach to generating referring expressions that we know of so far is the work of Poesio, Henschel, Hitzeman and Kibble (1999).[2] Their corpus is annotated with two types of factors:

1. factors that describe the NP to be generated, such as agreement information, semantic properties, and discourse factors,

2. factors that describe the antecedent, such as animacy, clause type, thematic role, and proximity

Their corpus consisted of descriptions of exhibitions furnished by museum guides. Poesio, Henschel, Hitzeman and Kibble (1999) trained CART trees (Breiman et al. 1984) on that corpus to predict surface forms of referring expressions. All 28 personal pronouns in their corpus were generated correctly. Unlike McCoy and Strube (1999), they do not evaluate the contribution of each of these factors.

Our detailed and dedicated feature selection experiments, as presented in sections 7.2 and 7.3, on the other hand, allow us to quantify the effect of each factor on the performance of the resulting algorithms and allow us to examine how these factors interact with each other, whether

[2]Machine Learning approaches to anaphora *resolution*, on the contrary, are far more numerous; recent examples are (Ge, Hale and Charniak 1998, Cardie and Wagstaff 1999, Connolly, Burger and Day 1997, Soon, Ng and Lim 1999). To discuss this work in more detail would lead us too far afield here.

they are complementary, or whether one factor, such as DIST, in fact subsumes others that have also received much attention in the literature, such as SYNANTE.

7.4.2 Potential Applications

The learning task we have considered here is quite artificial: given information about a referring expression, the discourse entity it refers to, and its antecedent, determine whether it should be pronominalised. Anaphora resolution is a completely different learning task: given a set of potential antecedents and a referring expression, determine whether that expression is anaphoric, and, if yes, which of the proposed antecedents is the real antecedent. First- and second person pronouns are easy to resolve if the potential addressees are well-defined and the text does not contain much direct speech. Therefore most anaphora resolution algorithms have focused on third person anaphora, more specifically on pronouns. For the purposes of anaphora resolution, we would need to recast the whole analysis in terms of selecting antecedents from a set of competing antecedents. The target variable PRO would need to be replaced by a variable called CORRECTANTECEDENT. In principle, we could then use the same methods that were used here for analysing that data and learning associations between anaphors and their antecedents.

Defining target variables such as PRO only makes sense from a generation perspective. In order to apply our approach to real-world generation systems, we will need to take into account the information made available to the system and the stage at which the form of the referring expression is generated. The values of AGREE and CLASS are determined by the entity that the referring expression accesses. Information about distance to last mention will be available in all systems that track when a discourse entity has last been mentioned. Additionally, the system will need to store information about the referring expression of that last mention. Our results indicate that information about its form (pronoun, definite, indefinite, demonstrative, other) might be sufficient. For calculating the number of competing antecedents, we need a slightly more sophisticated tracking mechanism which has access to the last 1-2 MCUs. The features SYN and PAR, which relies on SYN, presuppose that the syntactic role of the referring expression is known. Finally, we run into problems if the algorithm that determines constituent ordering does not come *before* the routine that calculates form of referring expression. In that case, we cannot determine the exact position of a referring expression in a co-specification sequence anymore, because we can never be sure whether there is an immediate antecedent in the same MCU. This problem affects the four antecedent-based features DIST, COMPANTE, FORMANTE and SYNANTE. Since these are the most powerful and robust predictors, it might be worthwhile to optimise the order of constituents and the form of referring expressions jointly. But if we do follow such a strategy, then we need to ask whether it still makes sense to investigate pronominalisation as a separate task. Although detailed, focused corpus studies such as those presented in this chapter provide necessary groundwork, the real fine-tuning comes when a real algorithm has to be integrated into a real system.

7.5 Summary of Main Results

The experiments in Section 7.2 and 7.3 have shown that distance to last mention predicts very well whether a discourse entity is to be referred to by a pronoun or not. The very simple

measure of entity status that I have defined and defended in Chapter 5.4 has proved to be more than adequate for large-scale corpus studies. These results are corroborated by the many studies on anaphora resolution and the form of referring expressions where the technical measure of distance played a key role (Ariel 1990, Ge et al. 1998, Mitkov 1998, to name but a few). In our search for factors that could supplement DIST, we restricted ourselves to factors that are easy to extract from existing annotations or that can be hand-coded reliably. Our motivation for this was not to develop a knowledge-poor approach to pronoun generation. Instead, it was theoretical: it does not make sense to run statistical tests on unreliable data if you are not prepared to make amends for inter-annotator variability. There are sophisticated techniques to deal with such noise, for example, dispersion parameters in the extended exponential family of probability distributions, or a subject variable in the statistical analysis, which ranges over the annotators that have contributed to a corpus. But to explore how these techniques further would have been beyond the scope of the thesis.

In this chapter, we explored the base case of pronominalisation: how can we account for *all* pronouns in twelve texts from diverse genres with no assumptions about hierarchical discourse structure? As far as I know, this is new; most previous studies have concentrated on third person pronouns, and cross-genre studies of the form of referring expressions are rare. Our results show that the influence of factors on pronominalisation varies greatly with genre. The only remotely robust factor is COMPANTE, the number of competing antecedents. Although we can identify a set of factors that perform well across genres, DIST, COMPANTE, FORMANTE, and AGREE, this combination does not always yield the best possible results. AGREE encodes the important distinction between animate (masculine, feminine) and inanimate (neuter) entities in the singular. The other three factors might reflect cognitive constraints on anaphora processing: recency (DIST), ambiguity (COMPANTE), and (discourse) topicality (FORMANTE).

Surprisingly, the detailed sortal class ontology was not very helpful. The most important distinctions were [$\pm$ animate] and [$\pm$ abstract]. The syntactic factors PAR, SYN, and SYNANTE were also not as important as researchers would suppose them to be. In our data, it was more important whether the antecedent is a pronoun than whether it acts as subject. We do not think that our classification of syntactic functions was too crude. In fact, our categories come very close to the hierarchy postulated by Givón (1992) for topicality, Subject $>$ Direct Object $>$ Other. The reason for the good performance of FORMANTE lies in the structure of co-specification sequences: When a discourse entity is in *active* mode, it tends to be pronominalised several times in a row. This pattern may be obscured by stylistic constraints, such as the repetition blocking that Henschel et al. (2000) observed in their data.

Let me close this empirical chapter with a theoretical note of caution: Corpus-based research is en vogue at the moment. More and more researchers are turning to corpora in order to replace hand-crafted algorithms by (hopefully) better automatically induced ones. But if some factors that experiments have shown to be important, such as thematic roles (Stevenson et al. 1994), cannot be annotated reliably in corpora (Poesio, Henschel, Hitzeman and Kibble 1999), then it may be much more effective to partition the task into problems that can be solved using Machine Learning, and components which still need to be hand-crafted.

8 Conclusion

In this chapter, I review the main results of the thesis, (Section 8.1), investigate the dimensions of givenness that I closed myself off from by focusing on the givenness of discourse entities, (Section 8.2), and come back full circle to the point I started out from: prosodic correlates of givenness in prosody (Section 8.3).

8.1 Main Results

8.1.1 Theoretical

Are there any theoretical results? Does a new theory of the givenness of discourse entities emerge from this thesis? No. I have drawn up a list of things that such a theory should cover; and the relevant information that has to be provided by this theory for each discourse entity is what I called *entity status*. Entity status is nothing special; similar catalogues of information must have been drawn up in many dissertations. The main difference here is that I shy away from devising a theory that provides this information. I prefer to keep my options open and explore competing theories, which all have something interesting to say in their own way. For example, generative grammar: How many parameters does it take to explain the form of referring expressions? Or Optimality Theory: How can present approaches to anaphora resolution be reformulated in terms of constraints (Beaver 2000)? Or Cognitive Grammar: How can we describe the constraints on pronominalisation in terms of conceptual reference points (van Hoek 1995, Langacker 1996)?

Entity status is a catalogue of demands which comes in two parts:

structural aspects: These can be characterised by three questions:

- In which segments does the discourse entity occur, and how are these segments connected?
- How is the discourse entity related to others in the discourse?
- How central is the entity? Is it part of the gist of the discourse (segment) it occurs in, or is it inconsequential?

management aspects: Three main functions can be identified:

- How are new discourse entities initialised? Where do addressees get sufficient information from to construct a good initial description, how can they best embed them in the current discourse model, when can they assumed that a potentially evoked discourse entity has been grounded, i.e. is available for referring back to?

- How are discourse entities accessed? Sheer salience, co-textual information, contextual information, world knowledge? How easy are they to access, and how easy is it to resolve to the wrong entity?
- How are representations of discourse entities updated? In particular, how do we handle changing properties, different points of view, how do we succeed in connecting a discourse entity that is central to the discourse even more tightly to the discourse model?

I probed some of the theories en vogue in linguistics and computational linguistics to see whether they could meet the list of demands of entity status. I found that all theories of discourse structure investigated fit the bill, in particular Rhetorical Structure Theory (Mann and Thompson 1988), which seems to have a veritable theory of structural entity status now with Veins Theory (Ide and Cristea 2000). The theory of Grosz and Sidner (1986) offers focus spaces, but there is still considerable debate about whether such a construct is needed (Walker 1996, Grosz and Gordon 1999). Finally, van Dijk's (1980) theory of macrostructures concentrates on how information is organised in texts, less on how discourse entities are maintained. However, it has the distinct advantage of being the basis of a very influential psycholinguistic theory of discourse comprehension, the Construction/Integration model outlined by Kintsch (1988) and its venerable predecessor, the theory of (van Dijk and Kintsch 1983).

For the theories that describe how discourse entities are managed, the picture is somewhat different. Talmy Givón's work is interesting because he uses corpus-based measures that are based on annotations which need to make only minimal assumptions about the current state of the speaker's nerves and the hearer's health, about the speaker's cooperativity and the hearer's basic stupidity. Just track the co-specification sequences there are and derive good counts, that is the basis of his strategies. Although his measures can be criticised heavily, they are a good start. His work has another crucial advantage: it is formulated terms of cognitive processing instructions that are compatible with one of the standard models of discourse comprehension, the Construction/Integration model. The theory can be tested on corpora, in simulations, in experiments. Givón may be wrong on some counts, but at least he is explicit enough so that one can prove him so. Another alluring approach is that of Wallace Chafe. The problem with his work—and its allure, paradoxically—is that it presents grave methodological problems. Chafe takes a fundamentally communicative perspective; he interprets his data in terms of what is supposed to go on in the consciousness of communicators and addressees. The difficulty with such analyses is that one quickly runs into problems of circularity. Take the word "consciousness". Was it semiactive when I mentioned it again? Did you quickly forget about that consciousness business while you processed the following sentence, eagerly waiting for some cutting remark on circularity? Or did the term "consciousness" stay at the top of your awareness, maybe because you are interested in neurolinguistics, maybe because you know Chafe's theories and wondered what I had to say about that term?

Why do I take this agnostic stance here? Why have I proved myself unable to commit to any single theory? Well, that is the fault of Appendix D. Those deeply rooted convictions that I do have come from a meta-linguistic, communication theoretic level. What leads me to prefer or disprefer a particular linguistic theory is not only whether it can explain the explicanda, but also whether it connects well with the communication theoretic ideas that I have become convinced of, and whether it is compatible with other theories that cover the same domain. For

the epistemological background of what I have just said, see (Schröder 1999). Of particular relevance to entity status are not only theories of communication, but also psycholinguistic theories of language production and understanding. The account that interfaces best with existing linguistic knowledge is, as far as I can see, Sanford and Garrod's Scenario Mapping and Focus Theory. My predilection for statistics as a method for describing patterns of language use also fits with Gerold Ungeheuer's general approach: Statistics is the ultimative extra-communicative way of looking at the system of language. Since both perspectives, communicative and extra-communicative, need to complement each other (Ungeheuer 1970/1972b), I have no problems with donning the hat of a formal statistician in one minute and exchanging it for the hat of the woolly discourse analyst in the next, as long as I do not overestimate the value of the results obtained either way.

8.1.2 Empirical

The empirical investigations of entity status reported in this thesis fell into three large parts: an exploration of how entity status can be studied in corpora (Chapter 5), a genre-specific study of entity status in radio news (Chapter 6), and a corpus-based study of influences on pronominalisation (Chapter 7). For these studies, I drew on different methods: in-depth analysis of the communication situation, interviews, corpus annotation, interpretation, statistical tests, statistical modelling techniques, and Machine Learning.

The radio news study showed clearly that givenness is almost impossible to define in the context of mass communication. Rather, it needs to be translated into categories that are more appropriate to the genre, such as news factors. A traditional linguistic approach using the well-known categories of Prince (1981) or Gundel et al. (1993) needs a very explicit addressee model and thus leaves great leeway to the analyst. The German and the American English corpora differ not only in the distribution of referring expressions, but also in the genre they belong to, which is in turn conditioned by the cultural differences in German and American radio. The German stories adhere mostly to the classic "lead-background-source" structure. This is what makes them coherent. In both corpora, definite descriptions appear to be the unmarked form for referring expressions. There is no time for telling stories in detail, there are few central referents. What is central is the discourse topic, and the goal is to cram all potentially relevant aspects of that topic into fifteen to a hundred seconds of speech. Accordingly, pronouns find their antecedents frequently in the same sentence, and indefinites tend to be used for conveying circumstantial information, not for introducing central discourse entities. Should linguists care about such results? Should they wipe them away with the remark that what I have studied is not "real" language? Well, I would most definitely not claim that I have shown that "in German texts, indefinites tend to etc.". Instead, I insist that for the moment at least, my results are restricted to one particular genre, and to be honest, to a sample from that genre that was mainly chosen according to the availability of prosodic annotations. However, I hypothesise that a replication of this study on a larger set of German radio news texts from DLF (now DeutschlandRadio), and on a larger set of correspondents' reports for news-oriented American public radio stations, would yield similar results. The main methodological conclusion from this study is that it can pay to take the genre one is analysing seriously. But because of the amount of work that this involves (for a really intimidating catalogue, see Bhatia 1993), it appears perfectly reasonable to abstract away from the communicative context as well—in particular if

results on a small, but well-analysed data set are to be substantiated by the analysis of a more comprehensive sample.

In this thesis, such a comprehensive sample was drawn from the Brown Corpus of American English (Francis and Kučera 1979), more precisely from these files that are also part of the Penn Treebank (LDC 1995). That sample, BROWN-COSPEC, is documented more fully in Appendix C. Since most of these texts are excerpts from longer texts, and since there is a certain temporal and cultural distance between texts from 1961 and a researcher that was born in 1974, a detailed analysis like that I performed on the radio news data becomes all but impossible. Instead, we need to resort to information that is relatively easy to add to arbitrary discourse and that can be annotated reliably. That information is information about co-specification sequences. Other linguistic semantic variables such as countability or genericity that were taken into account in the earlier study were dropped. Although they might be useful tools for exploring the syntax-semantics interface in English, they cannot be annotated reliably in corpora (Poesio, Henschel, Hitzeman and Kibble 1999). We decided to derive as much information as possible from the existing annotations in the Treebank (Strube and Wolters 2000). Logistic regression experiments and Machine Learning experiments showed:

1. Distance to last mention (a crude operationalisation of structural entity status) predicts pronominalisation extremely well.

2. Whether the antecedent is a pronoun influences the form of the anaphor more consistently than whether the antecedent is in subject position.

3. It is not possible, at least not with the factors we investigated, to develop a pronominalisation algorithm that performs well across genres. Whether discourse structure can redeem us is subject of future work.

Developing a statistical model of co-specification sequences on the basis of BROWN-COSPEC turned out to be very difficult. Although modeling the distribution of the mentions of a discourse entity in a text via a Poisson process gives us a rough approximation of the patterns we find, it cannot cover long-distance anaphora and constraints on the number of mentions within one unit. What we need is a non-stationary approach that distinguishes between different states of a discourse entity, maybe along the salience parameter that we identified in our discussion of the management aspects of entity status.

8.2 What about Givenness?

Givenness is a popular metaphor in linguistics; in some areas such as word order or intonational focus, it almost seems to have become a metaphor that linguists live by. But to explicate that metaphor, to turn it into a technical term, is very difficult. In the process, it loses much of its picturesque sweepingness. As Prince (1981) has already suggested, and as scholars such as Lambrecht (1994) have argued quite explicitly, the givenness of discourse entities should be kept strictly separate from the givenness of information, the givenness of the semantic content of a message. I followed that wise advice here and broke down the "givenness of discourse entities" into the prosaic facets of entity status. And when I searched prosody for correlates of that very specific type of givenness in a very specific genre, I found very few.

But what about the other givenness, the givenness of information? It is not monolithic, either. I would guess that this givenness can be described quite well by three dimensions well known from the literature (Halliday 1967, Kuno 1972, Clark and Haviland 1977), predictability, recoverability, and shared knowledge. *Shared knowledge* describes what all participants in the communication process know at the moment when they are communicating. It is "givenness in the present". *Predictability* describes what participants expect to happen, how participants expect the others to behave. What is predictable is boring; that was the result of experiments where subjects were asked to judge stories that conformed to Schank's (1977) scripts (Brewer and Lichtenstein 1981). Deviant stories were much more likely to be remembered than well-behaved ones. What is new is exciting, that catches the addressee's attention, that is what he has come to hear—if we assume that communication is mainly about exchanging information, that is. *Recoverability* describes the extent to which the utterance is phoric, the extent to which it needs to be interpreted with respect to the preceding co-text. In contrast to Prince (1981), who lumps both predictability and recoverability together, I separate them out because they describe different ways of looking at givenness as it develops in time. Predictability is cataphoric; given what I have seen so far, what is likely to come next? Recoverability is anaphoric: how does that which I am hearing now relate to what has come before?

What is shared by speaker and hearer obviously affects the form of referring expression they can use, but the most appropriate model of that "common ground" is in my view cognitive. The same goes for salience, which depends to a great extent on the background on the basis of which speaker and addressee construct their understanding of a text.

If we want to formalise predictability and recoverability, it makes sense to replace more woolly taxonomies of givenness by hard and fast measures based on information theory, as for example Pan has done (Pan and McKeown 1999, Pan and Hirschberg 2000). From the perspective of (formal) semantics, information may be new if it is not entailed by the context (Schwarzschild 1999), if it provides more specific information about a discourse marker (Kuhn 1996), if it is not presupposed, but asserted (Lambrecht 1994). All of these formalisations make slightly different predictions, and it would be interesting to compare them in domains other than those they normally compete in, intonational focus.

8.3 The Full Circle: Prosody

In the first drafts of this thesis, this section was a rag bag of all the things I intended to do, but probably would not get around to in time. What an overarching theme . . . everything you always wanted to know about givenness, but were afraid to ask. Instead, I have decided to come full circle back to where this excursion into givenness began. What have I learnt? What will I do differently now? Which questions would I like to ask in the future?

I would most definitely not study givenness on radio news any more. The communication situation is so complex (c.f. Section 6.1.3) that it is almost impossible to operationalise givenness in any cognitively satisfying way. What we can measure are genre-specific aspects: How are news factors (c.f. Section 6.1.2) verbalised? How do news readers phrase the lead sentence? Another question that comes to mind is: Looking at the sheer complexity of the sentences that radio news readers are faced with, then, how do they manage to convert them from something

designed for the medium of writing, where the addressee can go back over any lengthy hypotaxis he has not quite understood, to the medium of speaking, where a word is here now and gone in the next minute? Attacks such as those of Bolinger (1989) on radio news readers only highlight the problems involved in this task. From the functional point of view of e.g. Chafe (1994), the primary concern of these news readers would be to partition their text into information units that somehow conform to the "One New Idea Constraint". As to accentuation, text book writers such as Wachtel (1997) severely chide editors for putting new things first in a sentence. If news readers then want to emphasise that new information by a nuclear accent, they have to either fiddle with phrasing, or adopt a regular contour that sounds as if they were on Valium—an impression that is even encouraged by some professional German speakers when they train young news readers (Ralf Backhausen, personal communication), but not by others (Udo Stiehl, personal communication). There may also be cultural differences here; in America, getting the intonation right appears to be important (Margo Melnicove, personal communication). van Leeuwen (1984) suggests that the Valium effect might even be intended, because it suggests to the audience that the news reader is completely impartial. Thus, what we get in this genre, in particular in the up-market radio stations that the data analysed here comes from, is an interesting tradeoff between making your speech lively enough to be understood and sounding as neutral and distanced as necessary.

Future work on radio news prosody will therefore take me into two very different directions: The first direction will be to explore how listeners react to this typical news reading style, and to deviations from it. Will they really understand news better once they have been rewritten, once the phrasing cleanly separates chunks of information, once the core information is accented? And how will they react to that improved style? The second direction makes use of the fact that radio news texts can be very complex, and not at all adapted to reading them out loud. This makes radio news an ideal training ground for speech synthesis prosody modules. However, for training such modules, I would not apply any of the sophisticated approaches to entity status that I devised in Chapter 6. Rather, I would code co-specification sequences. Since there is more thematically homogeneous data in such texts, I would also compute co-occurrence probabilities and use straightforward information-theoretic measures of givenness, as suggested by Pan and McKeown (1999).

And now for something completely different . . .

Epilogue

A: What is my theory that it is? Yes.
Well, you may well ask what is my theory.
C: I am asking.
A: And well you may.
Yes, my word, you may well ask what it is, this theory of mine.
Well, this theory, that I have, that is to say, which is mine,
. . . is mine.
C: I know it’s yours! What is it?
A: . . . Where? . . . Oh! Oh! What is my theory?
C: Yes!
A: Ahh!
My theory, that I have, follows the lines that I am about to relate.
[starts prolonged throat clearing]
C: [under breath] Oh, God!
[Anne still clearing throat]
A: The Theory, by A. Elk (that’s ”A” for Anne”, it’s not by a elk.)
C: Right...
A: [clears throat] This theory, which belongs to me, is as follows...
[more throat clearing]
This is how it goes...
[clears throat]
The next thing that I am about to say is my theory.
[clears throat]
Ready?
C: [wimpers]
A: The Theory, by A. Elk [Miss].
My theory is along the following lines . . .
C: [under breath]God!
A: . . . All brontosauruses are thin at one end; much, much thicker in the middle and then thin again at the far end. That is the theory that I have and which is mine and what it is, too.
C: That’s it, is it?
A: Right, Chris!
C: Well, Anne, this theory of yours seems to have hit the nail right on the head.

from Monty Python’s Flying Circus, transcribed by `tim@zorac.arpa`

Bibliography

Aarts, J. (ed.): 1992, *New Directions in Corpus Linguistics*, Mouton, The Hague.

Aarts, J. and Meijs, W. (eds): 1990, *Theory and Practice in Corpus Linguistics*, Rodopi, Amsterdam.

Agresti, A.: 1990, *Categorical Data Analysis*, John Wiley, New York, NY.

Aha, D., Kibler, D. and Albert, M.: 1991, Instance-based learning algortihms, *Machine Learning* **6**, 37–66.

Akaike, H.: 1974, A new look at statistical model identification, *IEEE Trans. Automatic Control* **19**, 716–722.

Alexander, D.: 1961, *Bloodstain*, J.B. Lippincott, Philadelphia, PA. Excerpt from pages 128–134.

Altenberg, B.: 1992, Comment, *in* Aarts (1992), pp. 253–255.

Altrichter, M.: 1975, Das Rundfunknachrichtenmagazin - die News Show im Rundfunk, *in* Straßner (1975), pp. 242–254.

Ammann, H.: 1928, *Die menschliche Rede*, Moritz Schuenburg, Lahr i. B.

Andersen, E. B.: 1990, *The statistical analysis of categorical data*, Springer, New York, NY.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. and Weinert, R.: 1991, The HCRC Map Task Corpus, *Language and Speech* **34**(4), 351–366.

Antos, G., Brinker, K., Heinemann, W. and Sager, S. F. (eds): 2000, *Handbuch Text- und Gesprächslinguistik*, Vol. 1, de Gruyter, Berlin; New York, NY.

Aone, C. and Bennett, S. W.: 1995, Evaluating automated and manual acquisition of anaphora resolution strategies, *Proceedings of the 14th National Conference on Artificial Intelligence*, pp. 122–129.

Appelt, D. E.: 1985, Planning English referring expressions, *Artificial Intelligence* **26**(1), 1–33.

Ariel, M.: 1988, Retrieving propositions from context: why and how, *Journal of Pragmatics* **12**(3/4), 567–600.

Ariel, M.: 1990, *Accessing Noun Phrase Antecedents*, Routledge, London and New York, NY.

Ariel, M.: 1994, Interpreting anaphoric expressions: A cognitive versus a pragmatic approach, *Journal of Linguistics* **30**(1), 3–42.

Ariel, M.: 1998, The linguistic status of the "here and now", *Cognitive Linguistics* **9**(3), 189–238.

Arnold, J.: 1998, *Reference Form and Discourse Patterns*, PhD thesis, Stanford University.

Asher, N.: 1993, *Reference to Abstract Objects in Discourse*, Kluwer, Amsterdam.

Asher, N. and Lascarides, A.: 1998, Bridging, *Journal of Semantics* **15**(1), 83–113.

Baddeley, A.: 1998, *Human Memory*, revised edn, Allyn & Bacon, Needlam Heights, MA.

Badsberg, J.-H.: 1995, *An Environment for Graphical Models*, PhD thesis, Department of Mathematics and Computer Science, Aalborg University.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G. and Newlands, A.: 2000, Controlling the intelligibility of referring expressions in dialogue, *Journal of Memory and Learning* **42**(1), 1–22.

Bard, E. G., Robertson, D. and Sorace, A.: 1996, Magnitude estimation of linguistic acceptability, *Language* **72**(1), 32–68.

Barlow, S.: 1961, Monologue of murder, *The Saint Mystery Magazine* **15**(2), 121–125.

Bartlett, F.: 1932, *Remembering*, Cambridge University Press, Cambridge.

Barton, S. B. and Sanford, A. J.: 1993, A case-study of anomaly detection: Shallow semantic processing and cohesion establishment, *Memory and Cognition* **21**, 477–487.

Bátori, I. S., Lenders, W. and Putschke, W. (eds): 1989, *Computational Linguistics—Computerlinguistik*, Walter de Gruyter, Berlin.

Bäuerle, R., Schwarze, C. and v. Stechow, A. (eds): 1983, *Meaning, Use and Interpretation of Language*, de Gruyter, Berlin.

Beaver, D.: 2000, Centering in optimal theory. Department of Linguistics, Stanford University.

Beaver, D., Clark, B. Z. and Wolters, M.: in preparation, Is there a second-occurrence focus? Department of Linguistics, Stanford University.

Behrens, L.: 1995, Categorizing between lexicon and grammar. the MASS/COUNT distinction in a cross-linguistic perspective, *Lexicology* **1**, 1–112.

Behrens, L.: in preparation, Genericity from a cross-linguistic perspective. Institut für Sprachwissenschaft, Universität zu Köln.

Behrens, L. and Sasse, H.-J.: 1999, Qualities, Objects, Sorts, and Other Treasures: GOLD-digging in English and Arabic. Arbeitspapier Nr. 35 (Neue Folge), Institut für Sprachwissenschaft, Universität zu Köln.

Bell, A.: 1991, *Language in the News Media*, Blackwell, Oxford.

Bentele, G. and Rühl, M. (eds): 1993, *Theorien öffentlicher Kommunikation*, Ölschläger, München.

Bergenholtz, H. and Mugdan, J.: 1989, Korpusproblematik in der computerlinguistik, *in* Bátori, Lenders and Putschke (1989), pp. 141–150.

Berger, A. A.: 1995, *Essentials of Mass Communication Theory*, Sage, Thousand Oaks, CA.

Bergmann, P.: 2000, Zur theoretischen Konzeption von 'Topic' in Gesprächen. Master's Thesis, Universität Bonn.

Bhatia, V. K.: 1993, *Analysing Genre: Language Use in professional settings*, Longman, London.

Biber, D.: 1988, *Variation across Speech and Writing*, Cambridge University Press, Cambridge.

Biber, D.: 1992, Using computer–based text corpora to analyze the referential strategies of spoken and written texts, *in* Aarts (1992), pp. 213–252.

Bliss, E.: 1992, *Now the News: the Story of Broadcast Journalism*, Columbia University Press, New York, NY etc.

Bod, R. and Scha, R.: 1997, Data-oriented language processing, *in* Young and Bloothooft (1997), pp. 137–173.

Bolinger, D.: 1989, *Intonation and its Uses*, Arnold, London.

Boorstin, D. J.: 1961a, From news-gathering to news-making, *in* Schramm and Roberts (1971). originally in (Boorstin 1961b).

Boorstin, D. J.: 1961b, *The Image: A Guide to Pseudo-Events in America*, Atheneum.

Bortz, J.: 1993, *Statistik*, 4 edn, Springer, Berlin.

Bosch, P.: 1983, *Agreement and Anaphora: A Study of the Role of Pronouns in Syntax and Discourse*, Academic Press, London.

Botley, S. P.: 1996, Comparing demonstrative features in three written english genres, *in* Botley, McEnery and Wilson (1996), pp. 86–105.

Botley, S. P. and McEnery, A. M. (eds): 1998, *New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, Vol. 11 of *UCREL Technical Papers*, Lancaster University.

Botley, S. P., McEnery, A. M. and Wilson, A. (eds): 1996, *Approaches to Discourse Anaphora: Proceedings of the DAARC96 Colloquium*, Vol. 8 of *UCREL Technical Papers*, Lancaster University.

Brandt, M. and Rosengren, I.: 1992, Zur Illokutionsstruktur von Texten, *Zeitschrift für Literaturwissenschaft und Linguistik* **86**, 9–51.

Bransford, J., Barclay, J. and Franks, J.: 1972, Sentence memory: A constructive versus an interpretative approach, *Cognitive Psychology* **3**, 193–209.

Bratko, I. and Muggleton, S.: 1995, Applications of Inductive Logic Programming, *Communications of the ACM* **38**(11), 65–70.

Braunmüller, K.: 1977, *Referenz und Pronominalisierung*, Niemeyer, Tübingen.

Breheny, R.: 1997, A unitary approach to the interpretation of definites, *UCL Working Papers in Linguistics* **9**, 1–28.

Breiman, L., Friedman, J. H., Olshen, R. and Stone, C.: 1984, *Classification and Regression Trees*, Wadsworth and Brooks/Cole, Monterey.

Brennan, S. E.: 1995, Centering attention in discourse, *Language and Cognitive Processes* **10**(2), 137–167.

Brennan, S. E.: 1998, Centering as a psychological resource for achieving joint reference in spontaneous discourse, *in* Walker, Joshi and Prince (1998), pp. 227–249.

Brennan, S. E., Friedman, M. W. and Pollard, C. J.: 1987, A centering approach to pronouns, *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, CA, 6–9 July 1987, pp. 155–162.

Brewer, W. and Lichtenstein, E.: 1981, Event schemas, story schemas, and story grammars, *in* Long and Baddeley (1981).

Brewka, G. (ed.): 1996, *Advances in Knowledge Representation and Reasoning*, CSLI Publications, Stanford, CA.

Brinker, K.: 1997, *Linguistische Textanalyse*, 4 edn, Erich Schmidt, Berlin.

Brosius, H.-B.: 1990, Verstehbarkeit von fernsehnachrichten, *in* Wilke (1990), pp. 147–160.

Brown, C.: 1983a, Topic continuity in written english narrative, *in* Givón (1983c), pp. 313–342.

Brown, G.: 1983b, Prosodic structure and the Given/New distinction, *in* A. Cutler and D. R. Ladd (eds), *Prosody: Models and Measurements*, Springer, Berlin etc.

Brown, G.: 1995, *Speakers, Listeners and Communication*, paperback edn, Cambridge University Press, Cambridge.

Brown, G. and Yule, G.: 1983, *Discourse Analysis*, Cambridge University Press, Cambridge.

Brown, R. and Gilman, A.: 1960, The pronouns of power and solidarity, *American Anthropologist* **4**(6), 24–29. reprinted in (Fishman 1968, p. 252–275).

Bruce, R. and Wiebe, J.: 1999, Decomposable modeling in natural language processing, *Computational Linguistics* **25**(2), 195–207.

Bryson, L. (ed.): 1948, *The Communication of Ideas*, Institute for Religious and Social Studies, New York, NY.

Bucher, H.-J.: 1986, *Pressekommunikation*, Niemeyer, Tübingen.

Buchholz, A. and LaRoche, W. (eds): 1991, *Radio-Journalismus. Ein Handbuch für Ausbildung und Praxis im Hörfunk*, List, München.

Bühler, K.: 1927/1965, *Die Krise der Psychologie*, third, unchanged edn, Fischer, Stuttgart.

Bühler, K.: 1934, *Sprachtheorie*, Gustav Fischer, Jena. Reprint 1982.

Burger, H.: 1990, *Sprache der Massenmedien*, de Gruyter, Berlin.

Büring, D.: 1996, *The 59th Street Bridge Accent – On the Meaning of Topic and Focus –*, PhD thesis, Seminar für Sprachwissenschaft, Universität Tübingen. SFS-Report 05-96.

Büring, D. and Hartmann, K.: 1998, Asymmetrische Koordination, *Linguistische Berichte* **174**, 172–201.

Burnham, K. and Anderson, D.: 1998, *Model selection and inference: A practical information-theoretic approach*, Springer, New York, NY.

Bybee, J., Haiman, J. and Thompson, S. (eds): 1997, *Essays on language function and language type dedicated to Talmy Givón*, John Benjamins, Amsterdam.

Byron, D.: 1999, Resolving pronominal reference to abstract entities: Thesis proposal, *Technical Report 714*, Department of Computer Science, University of Rochester.

Cahn, J.: 1998, *A Computational Memory and Processing Model for Prosody*, PhD thesis, Massachusetts Institute of Technology Media Lab.

Cardie, C.: to appear, Integrating case-based learning and cognitive biases for machine learning of natural language, *J. Experimental and Theoretical Artificial Intelligence* .

Cardie, C. and Wagstaff, K.: 1999, Noun phrase coreference as clustering, *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Md., 21–22 June 1999, pp. 82–89.

Carletta, J.: 1996, Assessing agreement on classification tasks: The Kappa statistic, *Computational Linguistics* **22**(2), 249–254.

Carlson, G.: 1977, *Reference to Kinds in English*, PhD thesis, University of Massachussetts at Amherst.

Carlson, G.: 1991, Natural kinds and common nouns, *in* von Stechow and Wunderlich (1991), pp. 370–398.

Carlson, G. and Pelletier, F. (eds): 1995, *The Generic Book*, Chicago University Press, Chicago, IL.

Carreiras, M. and Gernsbacher, M. A.: 1992, Comprehending conceptual anaphors in Spanish, *Language and Cognitive Processes* **7**, 281–299.

Chafe, W.: 1976, Givenness, contrastiveness, definiteness, subjects, topics, and point of view, *in* Li (1976), pp. 25–55.

Chafe, W.: 1987, Cognitive constraints on information flow, *in* Tomlin (1987b), pp. 21–51.

Chafe, W.: 1994, *Discourse, Consciousness, and Time*, University of Chicago Press, Chicago, IL; London.

Chafe, W. (ed.): 1980, *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, Ablex, Norwood, NJ.

Charolles, M.: 1999, Associative anaphora and its interpretation, *Journal of Pragmatics* **31**(3), 311–326.

Chesterman, A.: 1998, *Contrastive Functional Analysis*, John Benjamins, Amsterdam.

Chomsky, N.: 1981, *Lectures on Government and Binding*, Foris, Dordrecht.

Church, K. W.: 2000, Empirical estimates of adaptation: The chance of two Noriegas is closer to p/2 than p^2, *Proceedings of the 18th International Conference on Computational Linguistics (COLING), Saarbrücken, Germany, 31 July - 4 August 2000*, pp. 180–186.

Churchill, C.: 1961, *A Notebook for the Wines of France*, Alfred A. Knopf, New York, NY. Excerpt from pages 124–129.

Cifuentes Honrubia, J. L. (ed.): 1998, *Estudios de Lingüística Cognitiva I*, Universidad de Alicante, Alicante.

Clancy, P. M.: 1980, Referential coherence in Japanese and English, *in* Chafe (1980), pp. 127–202.

Clancy, P. M.: 1992, Referential strategies in the narratives of Japanese children, *Discourse Processes* **15**, 441–467.

Clancy, P. M.: 1996, Referential strategies and the co-construction of argument structure in Korean acquisition, *in* Fox (1996), pp. 33–68.

Clark, H. H.: 1977, Bridging, *in* Johnson-Laird and Wason (1977), pp. 411–420.

Clark, H. H. and Haviland, S.: 1977, Comprehension and the given-new contract, *in* R. O. Freedle (ed.), *Discourse Production and Comprehension*, Ablex, pp. 1–40.

Clark, H. H. and Marshall, C. R.: 1981, Definite reference and mutual knowledge, *in* Joshi, Webber and Sag (1981), pp. 10–62.

Clark, H. H. and Wilkes-Gibbs, D.: 1990, Referring as a collaborative process, *in* P. Cohen, J. Morgan and M. Pollack (eds), *Intentions in Communication*, MIT Press, Cambridge, MA, pp. 463–494.

Coffin, T. P.: 1961, Folklore in the American Twentieth Century, *American Quarterly* **13**(4), 526–530.

Cohen, J.: 1960, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

Cohen, W.: 1995, Fast effective rule induction, *Machine Learning: Proceedings of the 12th International Conference Lake Tahoe, 1995*.

Cohen, W.: 1996, Learning trees and rules with set-valued features, *Proceedings of the 13th National Conference on Artificial Intelligence,* Seattle, Wash., 31 July – 4 August 1996.

Connolly, D., Burger, J. D. and Day, D. S.: 1997, A machine learning approach to anaphoric reference, *in* D. Jones and H. Somers (eds), *New Methods in Language Processing*, Oxford University Press, pp. 133–143.

Connolly, J.: 1989, *Philosophische Handlungstheorie*, Fernuniversität–Gesamthochschule Hagen, Hagen.

Cornish, F.: 1986, *Anaphoric Relations in English and French: A Discourse Perspective*, Croom Helm, London.

Corum, C.: 1973, Anaphoric peninsulas, *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*, pp. 89–97.

Coseriu, E.: 1975, *Die Geschichte der Sprachphilosophie von der Antike bis zur Gegenwart. Teil 1: Von der Antike bis Leibniz*, Gunter Narr, Tübingen.

Cote, S.: 1998, Ranking forward-looking centers, *in* Walker et al. (1998), pp. 55–69.

Cover, T. and Thomas, J.: 1991, *Elements of Information Theory*, Wiley-Interscience, New York, NY.

Cowart, W.: 1997, *Experimental Syntax*, Sage, Thousand Oaks, CA.

Cox, D. and Miller, H.: 1965, *The Theory of Stochastic Processes*, Methuen, London.

Cristea, D., Ide, N., Marcu, D. and Tablan, V.: 2000, An empirical investigation of the relation between discourse structure and co-reference, *Proceedings of the 18th International Conference on Computational Linguistics (COLING), Saarbrücken, Germany, 31 July - 4 August 2000*, pp. 208–214.

Cristea, D., Ide, N. and Romary, L.: 1998, Veins theory: A model of global discourse cohesion and coherence, *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Québec, Canada, 10–14 August 1998, Vol. 1, pp. 281–285.

Culicover, P. and McNally, L. (eds): 1998, *The Limits of Syntax*, Vol. 28 of *Syntax and Semantics*, Academic Press, New York, NY; London.

da Rocha, M. A. E.: 1997, Supporting anaphor resolution in dialogues with a corpus-based probabilistic model, *Proceedings of the ACL-97/EACL-97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text, Madrid, Spain, July 1997*.

da Rocha, M. A. E.: 1998, The antecedent likelihood heory: a methodology to analyse and resolve anaphora in dialogues, *Cognitive Science Research Papers 484*, School of Cognitive and Computing Sciences, University of Sussex.

Daelemans, W., van den Bosch, A. and Weijters, T.: 1997, IGTree: Using trees for compression and classification in lazy learning algorithms, *Artificial Intelligence Review* **11**, 407–423.

Daelemans, W., van den Bosch, A. and Zavřel, J.: 1999, Forgetting exceptions is harmful in language learning, *Machine Learning* **34**, 11–43.

Daelemans, W., Zavřel, J., van der Sloot, K. and van den Bosch, A.: 1999, TiMBL: Tilburg Memory Based Learner, *Technical Report ILK 99-01*, ILK Tilburg.

Dale, R. and Reiter, E.: 1995, Computational interpretations of the Gricean maxims in the generation of referring expressions, *Cognitive Science* **18**, 233–263.

Daneman, M. and Carpenter, P.: 1983, Individual differences in integrating information between and within sentences., *Journal of Experimental Psychology: Learning Memory and Cognition* **9**, 561–584.

Daneš, F.: 1974a, Functional sentence perspective and the organization of the text, in *Papers on Functional Sentence Perspective* (Daneš 1974b), pp. 106–128.

Daneš, F. (ed.): 1974b, *Papers on Functional Sentence Perspective*, Academia / Mouton, Prague / The Hague; Paris.

Darlington, R.: n.d., Measures of association in crosstab tables. `http://comp9.psych.cornell.edu/Darlington/crosstab/table0.htm`, last retrieved September 23, 2000.

Davidson, R., Schwarts, G. and Shapiro, D. (eds): 1986, *Consciousness and self-regulation. Advances in research and theory*, Plenum Press, New York, NY.

Davies, S. and Poesio, M.: 1998, Coding schemes for coreference. Chapter 3 in (*Supported Coding Schemes* 1998).

Davison, A.: 1984, Syntactic markedness and the definition of sentence topic, *Language* **60**, 797–846.

de Beaugrande, R.-A. and Dressler, W. U.: 1981, *Einführung in die Textlinguistik*, Niemeyer, Tübingen.

de Haan, P.: 1987, Exploring the linguistic database: Noun phrase complexity and language variation, *in* Meijs (1987), pp. 151–165.

de Haan, P.: 1992, The optimum corpus sample size?, *in* Leitner (1992), pp. 3–19.

de Moennink, I.: 1997, Using corpus and experimental data: a multimethod approach. `http://lands.let.kun.nl/literature/demonnink.1997.2.ps`.

de Mulder, W., Tasmowski-De Ryck, L. and Vetters, C. (eds): 1997, *Relations anaphoriques et (in)cohérence*, Rodopi, Amsterdam; Atlanta, GA.

De Renzi, E., Liotti, M. and Nichelli, N.: 1987, Semantic amnesia with preservation of autopbiographic memory. A case report, *Cortex* **23**, 575–597.

de Saussure, F.: 1916/1985, *Cours de linguistique générale*, Bibliothèque scientifique Payot, Paris, France.

DeCristofaro, J., Strube, M. and McCoy, K. F.: 1999, Building a tool for annotating reference in discourse, *Proceedings of the ACL '99 Workshop on the Relationship between Discourse/Dialogue Structure and Reference, University of Maryland, Maryland, 21 June, 1999*, pp. 54–62.

Dekker, P.: 1998, Speaker's reference, descriptions, and information structure, *Journal of Semantics* **15**(4).

Dewey, T. B.: 1961, *Hunter at Large*, Simon and Schuster, New York, NY. Excerpt from pages 118–124.

Di Eugenio, B.: 1998, Centering in Italian, *in* Walker et al. (1998), pp. 115–137.

Dickinson, C. and Givón, T.: 1997, Memory and conversation, *in* T. Givón (ed.), *Conversation*, John Benjamins, Amsterdam/Philadelpia, pp. 91–132.

Dik, S. C.: 1989, *The Theory of Functional Grammar. Part I: The Structure of the Clause*, Foris, Amsterdam.

Dirven, R. and Fried, V. (eds): 1987, *Functionalism in linguistics*, Mouton de Gruyter, The Hague.

Donnellan, K.: 1966, Reference and definite descriptions, *The Philosophical Review* **LXXV**, 281–304.

Donsbach, W. and Mathes, R.: 1999, Rundfunk, *in* Noelle-Neumann, Schulz and Wilke (1999), pp. 475–518.

Douloureux, P. T.: 1971, A note on one's privates, *in* Zwicky, Salus, Binnick and Vanek (1971), pp. 45–51.

Dressler, W.: 1974, Funktionelle Satzperspektive und Texttheorie, *in* Daneš (1974b), pp. 87–105.

Duda, R. O. and Hart, P. E.: 1973, *Pattern Classification and Scene Analysis*, Wiley, New York, NY.

EAGLES: 1996a, Preliminary recommendations on corpus typology, EAG—TCWG—CTYP/P, *Technical report*, Expert Advisory Group on Language Engineering Standards.

EAGLES: 1996b, Preliminary recommendations on text typology, EAG—TCWG—TTYP/P, *Technical report*, Expert Advisory Group on Language Engineering Standards.

Eckert, M. and Strube, M.: 1999, Resolving discourse deictic anaphora in dialogues, *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 8–12 June 1999, pp. 37–44.

Eckert, M. and Strube, M.: to appear, Dialogue acts, synchronising units and anaphora resolution, *Journal of Semantics* .

Eco, U.: 1994, *Einführung in die Semiotik*, Wilhelm Fink, Stuttgart. Italian original: *La struttura ausente*. 1968.

Eco, U.: 2000, *Kant und das Schnabeltier*, Hanser, München; Wien. Translated by Frank Herrmann. Italian original: *Kant e l'ornitorinco*. Mailand: Bompiani, 1997.

Ehlich, K.: 1982, Anaphora and deixis: Same, similar, or different?, *in* Jarvella and Klein (1982), pp. 315–338.

Eilders, C.: 1998, *Nachrichtenfaktoren und Rezeption: eine empirische Analyse zur Auswahl und Verarbeitung politischer Information*, Westdeutscher Verlag.

Eisenberg, P.: 1994, *Grundriß der deutschen Grammarik*, Metzler, Stuttgart.

Eisenhower, D. D.: 1961, *Peace with Justice*, Columbia University Press, New York, NY.

Elias, N.: 1970, *Einführung in die Soziologie*, Juventa, München.

Eysenck, M. W. and Keane, M. T.: 1995, *Cognitive Psychology*, 3 edn, Psychology Press, Hove.

Fanselow, G. and Felix, S.: 1987, *Sprachtheorie 2: Die Rektions- und Bindungstheorie*, Franke, Tübingen.

Fellbaum, C. (ed.): 1998, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Figueras Solanilla, C.: 1998, Semántica y pragmática de las expresiones anafóricas, *in* Cifuentes Honrubia (1998), pp. 61–76.

Firbas, J.: 1974, Some aspects of the czechoslovak approach o problems of functional sentence perspective, *in* Daneš (1974b), pp. 11–37.

Firbas, J.: 1992, *Functional sentence perspective in written and spoken communication*, Cambridge University Press, Cambridge.

Firth, J.: 1950, Personality and language in society, *Papers in Linguistics 1934–1957*, Oxford University Press, London, pp. 177–189.

Fishman, J. (ed.): 1968, *Readings in the Sociology of Language*, Mouton, The Hague.

Fletcher, C. R.: 1994, Levels of representation in memory for discourse, *in* Gernsbacher (1994), pp. 589–608.

Fligelstone, S.: 1992, Developing a scheme for annotating text to show anaphoric relations, *in* Leitner (1992), pp. 153–170.

Foertsch, J. and Gernsbacher, M. A.: 1994, In search of complete comprehension: Getting "minimalists" to work, *Discourse Processes* **18**, 271–296.

Fox, B.: 1987, *Discourse Structure and Anaphora*, Cambridge University Press, Cambridge.

Fox, B. (ed.): 1996, *Studies in Anaphora*, John Benjamins, Amsterdam/Philadelphia.

Francis, H. S., Gregory, M. and Michaelis, L.: 1998, Are lexical subjects deviant?, *Papers from the 34th Regional Meeting of the Chicago Linguistic Society*.

Francis, W. and Kučera, H.: 1979, *Brown Corpus Manual*, revised and amplified edn, Brown University. `http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM`.

Franz, A.: 1997, Independence assumptions considered harmful, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 7–12 July 1997, pp. 182–189.

Fraurud, K.: 1990, Definiteness and the processing of noun phrases in natural discourse, *Journal of Semantics* **7**, 395–433.

Fraurud, K.: 1996, Cognitive ontology and NP form, *in* Fretheim and Gundel (1996), pp. 65–87.

Frege, G.: 1892, On sense and reference, *in* Moore (1993), pp. 23–42. originally published in *Zeitschrift für Philosopie und philosophische Kritik*, vol. 100, 1892, 25-50; translation from Geach, Peter and Black, Max (editors), *Translations from the Philosophical Writings of Gottlob Frege*, Oxford: Blackwell 1952.

Fretheim, T. and Gundel, J. (eds): 1996, *Reference and Referent Accessibility*, Benjamins, Amsterdam.

Fries, P.: 1995, Themes, methods of development, and texts, *in* Hasan and Fries (1995b), pp. 317–359.

Fritz, G.: 1982, *Kohärenz. Grundfragen der linguistischen Kommunikationsanalyse*, Narr, Tübingen.

Früh, W.: 1980, *Lesen, Verstehen, Urteilen. Untersuchungen über den Zusammenhang von Textgestaltung und Textwirkung*, Alber, Freiburg/München.

Früh, W.: 1992a, Analyse sprachlicher Daten. Zur konvergenten Entwicklung "qualitativer" und "quantitativer" Methoden, *in* Hoffmeyer-Zlotnik (1992), pp. 59–89.

Früh, W.: 1992b, *Medienwirkungen: das dynamisch-transaktionale Modell*, Westdeutscher Verlag.

Früh, W.: 1994, *Realitätsvermittlung durch Massenmedien: die permanente Transformation der Wirklichkeit*, Westdeutscher Verlag, Opladen.

Früh, W.: 1998, *Inhaltsanalyse. Theorie und Praxis*, UVK-Medien, Konstanz.

Früh, W. and Schönbach, K.: 1982, Der dynamisch-transaktionale Ansatz. Ein neues Paradigma der Medienwirkungen, *Publizistik* **27**, 74–88. reproduced in (Früh 1992b, 23–40).

Fürnkranz, J. and Widmer, G.: 1994, Incremental reduced error pruning, *Machine Learning. Proceedings of the Eleventh Annual Machine Learning Conference, New Brunswick, New Jersey*, Morgan Kaufmann.

Galtung, J. and Ruge, M. H.: 1965, The structure of foreign news, *Journal of Peace Research* **2**(1), 64–91.

Gans, H. J.: 1980, *Deciding what's News*, Vintage, New York, NY.

Garnham, A.: 1996, The other side of mental models: Theories of language comprehension, *in* Oakhill and Garnham (1996), pp. 35–52.

Garnham, A. and Oakhill, J.: 1992, Discourse proessing and text representation from a "mental models" perspective, *Language and Cognitive Processes* **7**, 193–204.

Garnham, A., Oakhill, J. and Cruttenden, H.: 1992, The role of implicit causality and gender cue in the interpretation of pronouns, *Language and Cognitive Processes* **7**(3/4), 231–255.

Garnham, A., Traxler, M., Oakhill, J. and Gernsbacher, M. A.: 1996, The locus of implicit causality effects in comprehension, *Journal of Memory and Learning* **35**, 517–543.

Garrod, S. C. and Sanford, A. J.: 1994, Resolving sentences in a discourse context: How discourse representation affects language understanding, *in* Gernsbacher (1994), pp. 675–698.

Garrod, S. and Sanford, A.: 1989, Discourse models as interfaces between language and the spatial world, *Journal of Semantics* **6**, 147–160.

Garside, R., Fligelstone, S. and Botley, S.: 1997, Discourse annotaton: anaphoric relations in corpora, *in* Garside, Leech and McEnery (1997), pp. 66–84.

Garside, R., Leech, G. and McEnery, A. (eds): 1997, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Longman, London; New York, NY.

Ge, N., Hale, J. and Charniak, E.: 1998, A statistical approach to anaphora resolution, *Proceedings of the Sixth Workshop on Very Large Corpora*, Montréal, Canada, pp. 161–170.

Gernsbacher, M. A.: 1991, Comprehending conceptual anaphors, *Language and Cognitive Processes* **6**, 81–105.

Gernsbacher, M. A. (ed.): 1994, *Handbook of Psycholinguistics*, Academic Press, San Diego, CA etc.

Gernsbacher, M. A. and Givón, T. (eds): 1995, *Coherence in spontaneous text*, John Benjamins, Amsterdam/Philadelphia.

Gernsbacher, M. A., Hargreaves, D. and Beeman, M.: 1989, Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency, *Journal of Memory and Learning* **28**, 735–755.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. and Krüger, L.: 1989, *The Empire of Chance*, Cambridge University Press, Cambridge.

Ginzburg, J.: 1996, An update semantics for dialogue, *Proceedings of the First International Workshop on Computational Semantics*, Tilburg, Netherlands, January 13-15 1999, pp. 110–120.

Givón, T.: 1983a, Topic continuity in discourse: an introduction, *in* T. Givón (ed.), *Topic Continuity in Discourse*, John Benjamins, Amsterdam; Philadelphia, PA, pp. 1–42.

Givón, T.: 1983b, Topic continuity in spoken english, *in* T. Givón (ed.), *Topic Continuity in Discourse*, John Benjamins, Amsterdam; Philadelphia, PA, pp. 343–364.

Givón, T.: 1992, The grammar of referential coherence as mental processing instructions, *Linguistics* **30**.

Givón, T.: 1995a, Coherence in text, coherence in mind, *in* Gernsbacher and Givón (1995), pp. 59–115.

Givón, T.: 1995b, *Functionalism and Grammar*, John Benjamins, Amsterdam.

Givón, T. (ed.): 1983c, *Topic Continuity in Discourse*, John Benjamins, Amsterdam; Philadelphia, PA.

Glas, R.: 1975, Das LIMAS-Korpus, ein Textkorpus für die deutsche Gegenwartssprache, *Lingustische Berichte* **40**, 63–66.

Glenberg, A. M.: 1997, What memory is for, *Behavior and Brain Sciences* **20**, 1–19.

Glenberg, A. M., Kurley, P. and Langston, W. E.: 1994, Analogical processes in comprehension: Simulation of a mental model, *in* Gernsbacher (1994), pp. 609–640.

Glenberg, A., Meyer, M. and Lindem, K.: 1987, Mental models contribute to foregrounding during text comprehension, *Journal of Memory and Learning* **26**, 69–83.

Goldfarb, C.: 1990, *The SGML Book*, Clarendon Press, Oxford.

Gordon, P. C., Grosz, B. J. and Gilliom, L. A.: 1993, Pronouns, names, and the centering of attention in discourse, *Cognitive Science* **17**, 311–347.

Gordon, P. C. and Hendrick, R.: 1997, Intuitive knowledge of linguistic co-reference, *Cognition* **62**, 325–370.

Gordon, P. C. and Hendrick, R.: 1998, The representation and processing of coreference in discourse, *Cognitive Science* **22**(4), 389–424.

Goscinny, R. and Uderzo, A.: 1971/1974, *Die Trabantenstadt*, Ehapa, Stuttgart. Translated by Gudrun Penndorf; originial in French.

Graesser, A. C. and Kreuz, R. J.: 1993, A theory of inference generation during text comprehension, *Discourse Processes* **16**, 145–160.

Greenbaum, S. and Quirk, R.: 1970, *Elicitation Experiments in English Linguistic Studies in Usage and Attitude*, Longman, London.

Gregory, R.: 1981, *Mind in Science*, Cambridge University Press, Cambridge UK; London. cited after (Eco 2000).

Groenendijk, J., Janssen, T. and Stokhof, M. (eds): 1981, *Formal Methods in the Study of Language*, Mathematisch Centrum Tracts, Amsterdam.

Groenendijk, J., Stokhof, M. and Veltman, F.: 1996, Coreference and modality, *The Handbook of Contemporary Semantic Theory*, Blackwell, Oxford; Cambridge, MA, pp. 179–214.

Grosz, B. and Gordon, P. C.: 1999, Conceptions of limited attention and discourse focus, *Computational Linguistics* **25**(4), 617–624.

Grosz, B. J., Joshi, A. K. and Weinstein, S.: 1995, Centering: A framework for modeling the local coherence of discourse, *Computational Linguistics* **21**(2), 203–225.

Grosz, B. J. and Sidner, C. L.: 1986, Attention, intentions, and the structure of discourse, *Computational Linguistics* **12**, 175–204.

Guarino, N.: 1998, Some ontological principles for designing upper level lexical resources, *Proceedings of the First Conference on Language Resources and Evaluation*, Granada, Spain, 28-30 May 1998, Vol. I, pp. 527–534.

Gülich, E. and Raible, W.: 1977, *Linguistische Textmodelle*, Fink, München.

Gundel, J. K.: 1985, Shared knowledge and topicality, *Journal of Pragmatics* **9**(1), 83–97.

Gundel, J. K.: 1988, Universals of topic-comment structure, *in* M. Hammond, E. Moravcsik and J. Wirth (eds), *Studies in syntactic typology*, John Benjamins, Amsterdam, pp. 209–239.

Gundel, J. K., Hedberg, N. and Zacharski, R.: 1993, Cognitive status and the form of referring expressions in discourse, *Language* **69**, 274–307.

Haaß, C.: 1994, *Radionachrichten - öffentlich-rechtlich versus privat*, Fischer, München.

Haegeman, L.: 1994, *An Introduction to Government and Binding Theory*, Blackwell, Oxford; Cambridge, MA.

Hagen, L. v.: 1995, *Informationsqualität von Nachrichten: Meßmethoden und ihre Anwendung auf die Dienste von Nachrichtenagenturen*, Westdeutscher Verlag, Opladen.

Hahn, U., Markert, K. and Strube, M.: 1996, A conceptual reasoning approach to textual ellipsis, *Proceedings of the 12th European Conference on Artificial Intelligence*, Budapest, Hungary, 11–16 August 1996, pp. 572–576.

Halliday, M. A. K.: 1967, Notes on transitivity and theme in English. part ii, *Journal of Linguistics* pp. 299–244.

Halliday, M. A. K.: 1978, *Language as a Social Semiotic*, Arnold, London.

Halliday, M. A. K.: 1994, *An Introduction to Functional Grammar*, 2 edn, Arnold, London.

Halliday, M. A. K. and Hasan, R.: 1976, *Cohesion in English*, London: Longman.

Hanke, M.: 1984, *Der maieutische Dialog*, Rader, Aachen.

Harley, T.: 1995, *The Psychology of Language*, Psychology Press, Hove, UK.

Harweg, R.: 1979, *Pronomina und Textkonstitution*, 2 edn, Fink, München.

Hasan, R. and Fries, P.: 1995a, Reflections on subject and theme, *in* Hasan and Fries (1995b), pp. XIII–XLV.

Hasan, R. and Fries, P. (eds): 1995b, *On Subject and Theme*, Benjamins, Amsterdam/Philadelphia.

Hawkins, J. A.: 1978, *Definiteness and indefiniteness: A study in reference and grammaticality prediction*, Croon Helm, London.

Hawkins, J. A.: 1991, On (in)definite articles: implicatures and (un)grammaticality prediction, *Journal of Linguistics* **27**(2), 649–659.

Heim, I.: 1983, File Change Semantics and the familiarity theory of definiteness, *in* Bäuerle, Schwarze and v. Stechow (1983), pp. 164–189.

Heim, I. and Kratzer, A.: 1998, *Semantics in generative grammar*, Blackwell, Malden, MA; Oxford.

Heinemann, W. and Viehweger, D.: 1991, *Textlinguistik. Eine Einführung*, Niemeyer, Tübingen.

Hemphill, L.: 1989, Topic development, syntac, and social class, *Discourse Processes* **12**, 267–286.

Henschel, R., Cheng, H. and Poesio, M.: 2000, Pronominalization revisited, *Proceedings of the 18th International Conference on Computational Linguistics (COLING), Saarbrücken, Germany, 31 July - 4 August 2000*, pp. 306–312.

Heum, M.: 1975, Die Subjektivität der öffentlich-rechtlichen Nachrichten, *in* Straßner (1975).

Heuser, H.: 1993, *Lehrbuch der Analysis 1*, Teubner, Stuttgart.

Hirschberg, J.: 1993, Pitch accent in context: Predicting prominence from text., *Artificial Intelligence* **63**, 305–340.

Hirschberg, J. and Grosz, B.: 1992, Intonational features of local and global discourse structure, *Proceedings of the DARPA Workshop on Spoken Language Systems, Arden House, February 1992*, pp. 441–446.

Hirschman, L. and Chinchor, N.: 1997, MUC-7 coreference task definition, `http://www.muc.saic.com/proceedings/`.

Hirschman, L., Robinson, P., Burger, J. and Vilain, M.: 1998, Automating coreference: The role of annotated training data, *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

Hitzeman, J. and Poesio, M.: 1998, Long-distance pronominalisation and global focus, *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998.

Hobbs, J.: 1979, Coherence and coreference, *Cognitive Science* **2**, 67–90.

Hockey, B. A.: 1998, *The Interpretation and Realisation of Focus. An Experimental Investigation of Focus in English and Hungarian*, PhD thesis, University of Pennsylvania.

Hoffman, B.: 1995, *The Computational Analysis of the Syntax and Interpretation of "Free" Word Order in Turkish*, PhD thesis, University of Pennsylvania. published as IRCS Technical Report IRCS-95-17.

Hoffmann, L.: 2000, Anapher im text, *in* Antos, Brinker, Heinemann and Sager (2000), pp. 295–304.

Hoffmann-Riem, W.: 1985, Die Struktur des amerikanischen Rundfunkwesens und deren Regulierung, *in* Prokop (1985), pp. 139–227.

Hoffmeyer-Zlotnik, J. H. P. (ed.): 1992, *Analyse verbaler Daten. über den Umgang mit qualitativen Daten*, Westdeutscher Verlag, Opladen.

Holmes, D. I.: 1994, Authorship attribution, *Computing and the Humanities* **28**, 87–106.

Horacek, H. and Zock, M. (eds): 1993, *New Concepts in Natural Language Generation: Planning, Realization, and Systems*, Pinter, London.

Hörmann, H.: 1979, *Meinen und Verstehen*, Suhrkamp, Frankfurt am Main.

Hovy, E.: 1998, Combining and standardizing large-scale, practical ontologies for machine, *Proceedings of the First Conference on Language Resources and Evaluation*, Granada, Spain, 28-30 May 1998, Vol. I, pp. 535–544.

Huang, Y.: 1993, A neo-Gricean pragmatic theory of anaphora, *Journal of Linguistics* **27**, 301–335.

Hudson-D'Zmura, S. and Tanenhaus, M. K.: 1998, Assigning antecedents to ambiguous pronouns: The role of the center of attention as the default assignment, *in* Walker et al. (1998), pp. 199–226.

Ide, N. and Cristea, D.: 2000, A hierarchical account of reference resolution, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, 3–6 October 2000.

Ihaka, R. and Gentleman, R.: 1996, R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**, 299–314.

Isard, A., McKelvie, D., Mengel, A. and Baum Moller, M.: 2000, The MATE workbench annotation tool, a technical description, *Proceedings of the Second Conference on Language Resources and Evaluation*, Athens, Greece, pp. 1565–1570.

Jackendoff, R.: 1972, *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge, MA; London.

Jackendoff, R.: 1990, *Semantic Structures*, MIT Press, Cambridge, MA.

Jackendoff, R.: 1992, Parts and boundaries, *in* Levin and Pinker (1992), pp. 9–47. papers originally published in *Cognition*, 41(1-3), 1991.

Jacobs, J.: 1999, The dimensions of topic-comment. Handout of Talk presented at the Working Group on Adding and Omitting, Conference of the Deutsche Gesellschaft für Sprachwissenschaft, 1999, Konstanz.

Jacobs, J. (ed.): 1992, *Informationsstruktur und Grammatik*, Vol. 4 of *Linguistische Berichte, Sonderheft*, Westdeutscher Verlag, Opladen.

Jacobs, J., von Stechow, A. and Sternefeld, W. (eds): 1993, *Handbuch Syntax*, de Gruyter.

Jarvella, R. J. and Klein, W. (eds): 1982, *Speech, Place, and Action*, Wiley, Chichester.

Johansson, S., Atwell, E., Garside, R. and Leech, G.: 1986, *The Tagged LOB Corpus. User's Manual*, Norwegian Computing Center for the Humanities, Bergen.

Johnson-Laird, P.: 1983, *Mental Models. Towards a Cognitive Science of Language, Inference and Consciousness*, Cambridge University Press, Cambridge.

Johnson-Laird, P. N., Byrne, R. and Schaeken, W.: 1992, Propositional reasoning by model, *Psychological Review* **99**, 418–439.

Johnson-Laird, P. N., Byrne, R. and Tabossi, P.: 1989, Reasoning by model: The case of multiple quantification, *Psychological Review* **96**, 658–673.

Johnson-Laird, P. N. and Wason, P. C. (eds): 1977, *Thinking: Readings in Cognitive Science*, Cambridge University Press, Cambridge.

Jordan, P. W.: 2000, *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*, PhD thesis, University of Pittsburgh.

Joshi, A. K., Webber, B. L. and Sag, I. A. (eds): 1981, *Elements of Discourse Understanding*, Cambridge, MA: Cambridge University Press.

Juchem, J. G.: 1984, *Zeichentheorien*, Fernuniversität Hagen, Hagen.

Juchem, J. G.: 1989, *Konstruktion und Unterstellung*, Nodus, Münster.

Juchem, J. G.: 1998, *Kommunikationssemantik*, Nodus, Münster.

Jucker, A. H.: 1992, *Social Stylistics. Syntactic Variation in British Newspapers*, Mouton de Gruyter, Berlin; New York, NY.

Jucker, A. H.: 1996, News actor labelling in british newspapers, *Text* **16**(3), 373–390.

Just, M. A. and Carpenter, P. A.: 1992, A capacity theory of comprehension: Individual differences on working memory, *Psychological Review* **99**(1), 122–149.

Kameyama, M.: 1998, Intrasentential centering: A case study, *in* Walker et al. (1998), pp. 89–112.

Kamp, H.: 1981, A theory of truth and semantic representation, *in* Groenendijk, Janssen and Stokhof (1981), pp. 277–322.

Kamp, H. and Reyle, U.: 1993, *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, Kluwer, Dordrecht.

Karttunen, L.: 1976, Discourse referents, *in* McCawley (1976), pp. 363–386.

Kay, P. and McDaniel, C. K.: 1979, On the logic of variable rules, *Language in Society* **8**, 151–187.

Keller, F.: 1998, Gradient grammaticality as an effect of selective constraint re-ranking, *Papers from the 34th Regional Meeting of the Chicago Linguistic Society*, pp. 95–109.

Kellerman, K., Broetzmann, S., Lim, T.-S. and Kitao, K.: 1989, The conversation MOP: Scenes in the stream of discourse, *Discourse Processes* **12**, 27–61.

Kempson, R.: 1988, Grammar and conversational principles, *in* Newmeyer (1988), pp. 139–163.

Kepplinger, H. M.: 1990, Realität, Realitätsdarstellung und Medienwirkung, *in* Wilke (1990), pp. 39–56.

Kepplinger, H. M.: 1993, Erkenntnistheorie und Forschungspraxis des Konstruktivismus, *in* Bentele and Rühl (1993), pp. 118–125.

Kepplinger, H. M.: 1999, Kommunikationspolitik, *in* Noelle-Neumann et al. (1999), pp. 116–140.

Kessler, B., Nunberg, G. and Schütze, H.: 1997, Automatic detection of text genre, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 7–12 July 1997, pp. 32–38.

Keysar, B.: 1997, Unconfounding common ground, *Discourse Processes* **24**, 253–270.

Kibble, R.: 1999, Cb or not Cb? Centering theory applied to NLG, *Proceedings of the ACL '99 Workshop on the Relationship between Discourse/Dialogue Structure and Reference, University of Maryland, Maryland, 21 June, 1999*.

Kibble, R. and Powers, R.: 1999, Using centering theory to plan coherent texts, *Proceedings of the 12th Amsterdam Colloqium, Amsterdam, December 1999*.

Kibble, R. and van Deemter, K.: 1999a, What is coreference, and what should coreference annotation be?, *Proceedings of the ACL '99 Workshop on Coreference and its Applications, University of Maryland, Maryland, June, 1999*.

Kibble, R. and van Deemter, K.: 2000, Coreference annotation: Whither?, *Proceedings of the Second Conference on Language Resources and Evaluation*, Athens, Greece, pp. 1281–1286.

Kibble, R. and van Deemter, K. (eds): 1999b, *Proceedings of the Workshop on The Generation of Nominal Expressions, 11th European Summer School on Logic, Language, and Information, Utrecht, 9-13 August 1999*.

Kibrik, A. A.: 1996, Anaphora in russian narrative prose, *in* Fox (1996), pp. 255–303.

Kilgariff, A.: 1998, Gold standard data sets for word sense disambiguation, *Computer Speech and Language* **12**(3), 453–472.

Kintsch, W.: 1985, Text processing: a psychological model, *in* van Dijk (1985c), pp. 231–244.

Kintsch, W.: 1988, The role of knowledge in discourse comprehension: a construction-integration approach, *Psychological Review* **95**, 163–182.

Kintsch, W.: 1993, Information accretion and reduction in text processing: Inferences, *Discourse Processes* **16**, 193–202.

Kintsch, W.: 1994, The psychology of text comprehension, *in* Gernsbacher (1994), pp. 721–740.

Kintsch, W.: 1995, How readers consruct situation models for stories: The role of syntactic cues and causal inferences, *in* Gernsbacher and Givón (1995), pp. 139–160.

Kintsch, W. and van Dijk, T. A.: 1978, Toward a model of text comprehension and production, *Psychological Review* **85**, 363–394.

Kleiber, G.: 1997, Anaphore nominale et référents évolutifs ou comment faire recette avec un pronom, *in* de Mulder, Tasmowski-De Ryck and Vetters (1997), pp. 1–30.

Klein, W. and von Stutterheim, C.: 1992, Textstruktur und referentielle Bewegung, *Zeitschrift für Literaturwissenschaft und Linguistik* **86**, 67–92.

Kniffka, H.: 1980, *Soziolinguistik und empirische Textanalyse: Schlagzeilen– und Leadformulierung in amerikanischen Tageszeitungen*, Niemeyer, Tübingen.

Knott, A.: 1996, *A Data-Driven Methodology for Motivating a Set of Coherence Relations*, PhD thesis, University of Edinburgh.

Knott, A. and Sanders, T. J.: 1998, The classification of coherence relations and their linguistic markers: An exploration of two languages, *Journal of Pragmatics* **30**(2), 135–172.

Kohavi, R. and John, G.: 1998, The wrapper approach, *in* H. Liu and H. Motoda (eds), *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer, Dordrecht.

Komegata, N.: 1999, *A computational analysis of informatin structure using parallel expository texts in English and Japanese*, PhD thesis, University of Pennsylvania.

Krahmer, E. and Theune, M.: 1999, Generating descriptions in context, *in* Kibble and van Deemter (1999b).

Krifka, M.: 1991, Massennomina, *in* von Stechow and Wunderlich (1991), pp. 399–417.

Krifka, M.: 1992, A compositional semantics for multiple focus constructions, *in* Jacobs (1992), pp. 17–53.

Krifka, M., Pelletier, F., Carlson, G., ter Meulen, A., Link, G. and Chierchia, G.: 1995, Genericity: an introduction, *in* Carlson and Pelletier (1995), pp. 1–124.

Krippendorff, K.: 1980, *Content Analysis*, Sage, Thousand Oaks, CA.

Krippendorff, K.: 1993, Schritte zu einer konstruktivistischen Erkenntnistheorie, *in* Bentele and Rühl (1993), pp. 19–51.

Krippendorff, K.: 1994, Der verschwundene bote. metaphern und modelle der kommunikation, *in* Merten et al. (1994), pp. 79–113.

Kruijff-Korbayová, I.: 1998, *The Dynamic Potential of Topic and Focus*, PhD thesis, Department of Mathematics and Physics, Charles University, Prague.

Kucera, H. and Francis, W.: 1967, *Frequency Analysis of English usage: Lexicon and Grammar*, Houghton Mifflin, Boston, MA.

Kuhn, J.: 1996, On intonation amd interpretation in context - is there a unitary explanation for focus and deaccenting? Master's Thesis, Universität Stuttgart.

Kuno, S.: 1972, Functional sentence perspective, *Linguistic Inquiry* **3**(3), 269–320.

Lambrecht, K.: 1994, *Information Structure and Sentence Form*, Cambridge University Press, Cambridge.

Langacker, R.: 1996, Conceptual grouping and pronominal anaphora, *in* Fox (1996), pp. 333–378.

Langley, P. and Simon, H. A.: 1995, Applications of machine learning and rule induction, *Communications of the ACM* **38**, 55–64.

Lappin, S. and Leass, H. J.: 1994, An algorithm for pronominal anphora resolution, *Computational Linguistics* **20**(4), 535–561.

LaRoche, W. v.: 1991a, Fürs Hören schreiben, *in* Buchholz and LaRoche (1991), pp. 54–67.

LaRoche, W. v.: 1991b, Nachrichten-präsentation, *in* Buchholz and LaRoche (1991), pp. 87–96.

Lasswell, H. D.: 1948, The structure and function of communication in society, *in* Schramm and Roberts (1971), pp. 84–99. originally in: (Bryson 1948).

Lazarsfeld, P. F., Berelson, B. and Gaudet, H.: 1944, *The People's Choice. How the Voter Makes up his Mind in a Presidential Campaign*, Duell, Sloan and Pearce, New York, NY.

Lazarsfeld, P. F. and Merton, R. K.: 1948/1971, Mass communication, popular taste, and organized social action, *in* Schramm and Roberts (1971), pp. 554–578. originally in (Bryson 1948).

LDC: 1995, Penn Treebank-II, Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.

Lee, D. Y.: submitted, Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. Lancaster University.

Leitner, G. (ed.): 1992, *New directions in English Language Corpora*, Mouton de Gruyter, Berlin.

Lenat, D.: 1995, Cyc: A large-scale investment in knowledge infrastructure, *Communications of the ACM* **38**(11).

Leopold, E.: 1998, *Stochastische Modellierung lexikalischer Evolutionsprozesse*, Dr. Kovač, Hamburg.

Levin, B. and Pinker, S. (eds): 1992, *Lexical and Conceptual Semantics*, Cognition Special Issues, Blackwell, Cambridge, MA; Oxford. papers originally published in *Cognition*, 41(1-3), 1991.

Levinson, S. C.: 1987, Pragmatics and the grammar of anaphora: a partial pragmatic reduction of binding and control phenomena, *Journal of Linguistics* **23**, 379–434.

Levinson, S. C.: 1991, Pragmatic reduction of the binding conditions revisited, *Journal of Linguistics* **27**, 107–162.

Levy, E.: 1982, Towards an objective definition of 'discourse topic', *Papers from the 18th Regional Meeting of the Chicago Linguistic Society*, pp. 295–304.

Li, C. N.: 1997, On zero anaphora, *in* Bybee, Haiman and Thompson (1997), pp. 275–300.

Li, C. N. (ed.): 1976, *Subject and Topic*, Academic Press, New York, NY.

Light, L. L., Capps, J. L., Singh, A. and Albertson Owens, S. A.: 1994, Comprehension and use of anaphoric devices in young and older adults, *Discourse Processes* **18**, 77–103.

Lindsay, J.: 1995, *Introducing Statistics. A Modellign Approach*, Oxford University Press, Oxford.

Linke, A., Nussbaumer, M. and Portmann, P. T.: 1994, *Studienbuch Linguistik*, Niemeyer, Tübingen.

Lippmann, W.: 1922, *Public Opinion*, Macmillan, New York, NY.

Lippmann, W.: 1922/1971, The world outside and the pictures in our heads, *in* Schramm and Roberts (1971), pp. 265–286. Chapter 1 from (Lippmann 1922).

Löbner, S.: 1985, Definites, *Journal of Semantics* **4**, 279–326.

Long, J. and Baddeley, A. (eds): 1981, *Attention and Performance*, Vol. IX, Erlbaum, Hillsdale, NJ.

Lötscher, A.: 1987, *Text und Thema: Studien zur thematischen Konstituenz von Texten*, Niemeyer, Tübingen.

Luckmann, T.: 1988, *Grundlagen der Soziologie: Strukturen sozialen Handelns*, Fernuniversität–Gesamthochschule Hagen, Hagen.

Lüger, H.-H.: 1983, *Pressesprache*, Niemeyer, Tübingen.

Lyons, C.: 1999, *Definiteness*, Cambridge University Press, Cambridge.

Maier, E. and Hovy, E.: 1993, Organizing discourse structure relations using metafunctions, *in* Horacek and Zock (1993), pp. 69–86.

Malinowski, B.: 1923, The problem of meaning in primitive languages, *in* C. K. Ogden and I. A. Richards (eds), *The meaning in of meaning*, Routledge, London.

Mani, I. and Maybury, M. (eds): 1999, *Advances in Text Summarization*, MIT Press, Cambridge, MA.

Mann, W. C., Matthiessen, C. M. and Thompson, S. A.: 1992, Rhetorical structure theory and text analysis, *in* Mann and Thompson (1992), pp. 39–78.

Mann, W. C. and Thompson, S. A.: 1987, Rhetorical structure theory: A framework for the analysis of texts, *IPrA Papers in Pragmatics* **1**, 79–105. also available as: Technical Report ISI/RS-87-185, USC Information Sciences Institute.

Mann, W. C. and Thompson, S. A.: 1988, Rhetorical structure theory. Toward a functional theory of text organization, *Text* **8**(3), 243–281.

Mann, W. C. and Thompson, S. A. (eds): 1992, *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*, John Benjamins, Amsterdam.

Marcu, D.: 1997, *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, PhD thesis, Department of Computer Science, University of Toronto.

Marcu, D.: 1999, Discourse trees are good indicators of importance in text, *in* Mani and Maybury (1999), pp. 123–136.

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A.: 1993, Building a large annotated corpus of English: The Penn Treebank, *Computational Linguistics* **19**, 313–330.

Marslen-Wilson, W., Levy, E. and Komisarjevsky Tyler, L.: 1982, Producing interpretable discourse: The establishment and maintenance of reference, *in* Jarvella and Klein (1982), pp. 339–378.

Mathesius, V.: 1929, Zur Satzperspektive im modernen Englisch, *Archiv für das Studium der neueren Sprachen und Literaturen* **155**, 202–210.

McCawley, J. D. (ed.): 1976, *Notes from the Linguistic Underground*, Vol. 7 of *Syntax and Semantics*, Academic Press, New York, NY; San Francisco, CA; London.

McCoy, K. F. and Strube, M.: 1999, Taking time to structure discourse: Pronoun generation beyond accessibility, *Proceedings of the 21st Annual Conference of the Cognitive Science Society*, Vancouver, British Columbia, Canada, 19–21 August 1999, pp. 378–383.

McCullagh, P. and Nelder, J.: 1983, *Generalized Linear Models*, Chapman and Hall, London.

McKoon, G. and Ratcliff, R.: 1992, Inference during reading, *Psychological Review* **99**(1), 440–466.

Meijs, W. (ed.): 1987, *Corpus Linguistics and Beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*, Rodopi, Amsterdam.

Mengel, A., Dybkjaer, L., Garrido, J. M., Heid, U., Klein, M., Pirelli, V., Poesio, M., Quazza, S., Schiffrin, M. and Soria, C.: 2000, Mate dialogue annotation guidelines, *Technical report*, MATE.

Merten, K.: 1994, Wirkungen von Kommunikation, *in* Merten et al. (1994), pp. 291–328.

Merten, K., Schmidt, S. J. and Weischenberg, S. (eds): 1994, *Die Wirklichkeit der Medien*, Westdeutscher Verlag, Opladen.

Meyn, H.: 1999, *Massenmedien in Deutschland*, UVK Medien, München.

Miller, A.: 1961a, The prophecy, *in* M. Foley and D. Burnett (eds), *The Best American Short Stories*, Houghton Mifflin, Boston, MA, pp. 258–262.

Miller, A. S.: 1961b, Toward a concept of national responsibility, *The Yale Review* **LI**, 186–191.

Miller, G. A.: 1995, Wordnet: A lexical database for English, *Communications of the ACM* **38**(11), 39–41.

Miller, G. and Fellbaum, C.: 1992, Semantic networks of english, *in* Levin and Pinker (1992), pp. 197–230. papers originally published in *Cognition*, 41(1-3), 1991.

Minsky, M.: 1975, A framework for representing knowledge, *in* Winston (1975), pp. 211–277.

Mitkov, R.: 1998, Robust pronoun resolution with limited knowledge, *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Québec, Canada, 10–14 August 1998, pp. 869–875.

Mixdorff, H.: 2000, A novel approach to the fully automatic extraction of fujisaki model parameters, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, 2000*, Vol. 3, pp. 1281—1284.

Mixdorff, H. and Fujisaki, H.: 2000, A quantitative description of German prosody offering symbolic labels as a by-product, *Proceedings of the International Conference on Spoken Language Processing, Beijing, China*.

Möhler, G.: 1998, Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese, *Arbeitspapiere des Instituts für maschinelle Sprachverabeitung, Stuttgart* **4**(2).

Molnár, V.: 1993, Zur Pragmatik und Grammatik des TOPIK-Begriffs, *in* M. Reis (ed.), *Wortstellung und Informationsstruktur*, Niemeyer, Tübingen, pp. 155–202.

Moore, A. (ed.): 1993, *Meaning and Reference*, Oxford University Press, Oxford.

Moore, J. D. and Pollack, M. E.: 1993, A problem for RST: The need for multi-level discourse analysis, *Computational Linguistics* pp. 537–544.

Moser, M. and Moore, J. D.: 1996, Toward a synthesis of two accounts of discourse structure, *Computational Linguistics* **22**(3), 409–419.

Mosteller, F. and Wallace, D.: 1964, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Cambridge, MA.

Motsch, W., Reis, M. and Rosengren, I.: 1990, Zum Verhältnis von Satz und Text, *Deutsche Sprache* (2), 97–125.

Muggleton, S. and De Raedt, L.: 1994, Inductive logic programming: Theory and methods, *Journal of Logic Programming* **19/20**, 629–679.

Mühlhäusler, P. and Harré, R.: 1990, *Pronouns and People*, Blackwell, Oxford; Cambridge, MA.

Müller, K. E.: 1998, German focus particles and their influence on intonation. Master's Thesis, Universität Stuttgart.

Murphy, G. L.: 1992, Comphrehension and memory of personal reference: The use of social information in language processing, *Discourse Processes* **15**, 337–356.

Myhill, J.: 1992, *Typological Discourse Analysis*, Blackwell, Cambridge, MA; Oxford.

Nakatani, C.: 1997, *The Computational Processing of Intonational Prominence: A Functional Prosody Perspective*, PhD thesis, Harvard University.

Newen, A.: 1995, *Kontext, Referenz und Bedeutung*, Schöningh, Paderborn.

Newmeyer, F. J. (ed.): 1988, *Linguistics: The Cambridge Survey*, Vol. 2: Linguistic Theory: Extensions and Implications, Cambridge University Press, Cambridge; New York, NY; Melbourne.

Noelle-Neumann, E.: 1999, Wirkung der Massenmedien auf die Meinungsbildung, *in* Noelle-Neumann et al. (1999), pp. 518–571.

Noelle-Neumann, E., Schulz, W. and Wilke, J. (eds): 1999, *Fischer-Lexikon Publizistik Massenkommunikation*, Fischer Taschenbuch Verlag.

Norman, D. A. and Rumelhart, D. E. (eds): 1978, *Strukturen des Wissens*, Klett-Cotta, Stuttgart. Translated by Urs Aeschbacher, Walter F. Bischof and Roman Müller. English Original: *Explorations in Cognition*, San Francisco, CA: Freeman, 1975.

Norman, D. A. and Shallice, T.: 1986, Attention to action: Willed and automatic control of behavior, *in* Davidson, Schwarts and Shapiro (1986), pp. 1–18.

Nussbaumer, M.: 1991, *Was Texte sind und was sie sein sollen*, Niemeyer, Tübingen.

Oakes, M. P.: 1998, *Statistics for Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Oakhill, J.: 1996, Mental models in children's text comprehension, *in* Oakhill and Garnham (1996), pp. 77–94.

Oakhill, J. and Garnham, A.: 1992, Linguistic prescriptions and anaphoric reality, *Text* **12**, 161–182.

Oakhill, J. and Garnham, A. (eds): 1996, *Mental Models in Cognitive Science. Essays in Honour of Phil Johnson-Laird*, Psychology Press, Hove.

Oakhill, J., Garnham, A., Gernsbacher, M. A. and Cain, K.: 1992, How natural are conceptual anaphors?, *Language and Cognitive Processes* **7**, 257–280.

Oppenheimer, F.: 1961, Science and fear — a discussion of some fruits of scientific understanding, *The Centennial Review* **5**(4), 404–409.

Ostendorf, M., Price, P. and Shattuck-Hufnagel, S.: 1995, The Boston University radio news corpus, *Technical report*, Boston University.

Pan, S.: 1998, Learning intonation rules for concept-to-speech generation. Unpublished Dissertation Proposal, Computer Science Department, Columbia University.

Pan, S. and Hirschberg, J.: 2000, Modeling local context for pitch accent prediction, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, 3–6 October 2000.

Pan, S. and McKeown, K.: 1999, Word informativeness and automatic pitch accent modelling, *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Md., 21–22 June 1999.

Passonneau, R.: 1997, Summary of the coreference group, *in* J. Carletta, N. Dahlbäck, N. Reithinger and M. A. Walker (eds), *Standards for Dialogue Coding in Natural Language Processing*, Discourse Resource Initiative, http://www.dfki.de/dri/, pp. 12–21.

Passonneau, R. J.: 1996, Instructions for applying discourse reference annotation for multiple applications (DRAMA). Columbia University, New York, Dept. of Computer Science.

Passonneau, R. J.: 1998, Interaction of discourse structure with explicitness of discourse anaphoric noun phrases, *in* Walker et al. (1998), pp. 327–358.

Passonneau, R. J. and Litman, D. J.: 1993, Intention based segmentation: Human reliability and correlation with linguistic cues, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 22–26 June 1993, pp. 148–155.

Passonneau, R. J. and Litman, D. J.: 1997, Discourse segmentation by human and automated means, *Computational Linguistics* **23**(1), 103–139.

Paul, H.: 1920, *Prinzipien der Sprachgeschichte*, 5 edn, Max Niemeyer, Halle an der Saale.

Pazzani, M. and Kibler, D.: 1992, The utility of knowledge in inductive learning, *Machine Learning* **9**(1), 57–94.

Poesio, M.: 2000, Coreference. Section 2.4 in (Mengel et al. 2000).

Poesio, M., Bruneseaux, F. and Romary, L.: 1999, The MATE meta-scheme for co-reference in dialogues in multiple languages, *Proceedings of the ACL '99 Workshop on Standards for Discourse Tagging, University of Maryland, Maryland, June, 1999*.

Poesio, M., Henschel, R., Hitzeman, J. and Kibble, R.: 1999, Statistical NP generation: A first report, *in* Kibble and van Deemter (1999b).

Poesio, M. and Traum, D.: 1997, Conversational actions and discourse situations, *Computational Intelligence* **13**(3), 309–347.

Poesio, M. and Vieira, R.: 1998, A corpus-based investigation of definite description use, *Computational Linguistics* **24**(2), 183–216.

Polanyi, L.: 1988, A formal model of the structure of discourse, *Journal of Pragmatics* **12**, 601–638.

Postal, P. M.: 1969, Anaphoric islands, *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, pp. 205–239.

Pratchett, T.: 1990, *Guards! Guards!*, Transworld Publishers, London; Moorebank; Auckland.

Prince, A. and Smolensky, P.: 1993, Optimality theory: Constraint interaction in generative grammar, *Technical Report 2*, Rutgers University Center for Cognitive Science RuCCS.

Prince, E. F.: 1981, Towards a taxonomy of given-new information, *in* P. Cole (ed.), *Radical Pragmatics*, Academic Press, New York, N.Y., pp. 223–255.

Prince, E. F.: 1992, The ZPG letter: Subjects, definiteness, and information-status, *in* W. Mann and S. Thompson (eds), *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*, John Benjamins, Amsterdam, pp. 295–325.

Prokop, D. (ed.): 1985, *Medienforschung. Konzerne, Macher, Kontrolleure*, Fischer, München.

Pu, M.-M.: 1995, Anaphoric patterning in English and Mandarin narrative production, *Discourse Processes* **19**, 279–300.

Pürer, H.: 1998, *Einführung in die Publizistikwissenschaft*, 6 edn, Ölschläger im UVK Medien, Konstanz.

Quinlan, J. R.: 1990, Learning logical definitions from relations, *Machine Learning* **5**(3), 239–266.

Quinlan, J. R.: 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufman, San Mateo, CA.

Quinlan, J. R.: 1996, Improved use of continuous attributes in C4.5, *Journal for Artificial Intelligence Reasearch* **4**, 77–90.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J.: 1985, *A Comprehensive Grammar of the English Language*, Longman, London.

Rabiner, L. and Juang, B.-H.: 1993, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ.

Radin, E. D.: 1961, *Lizzie Borden: The Untold Story*, Simon and Schuster, New York, NY. Excerpt from pages 208–214.

Rapp, S.: 1998, Automatisierte Erstellung von Korpora für die Prosodieforschung, *Arbeitspapiere des Instituts für maschinelle Sprachverabeitung, Stuttgart 1*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Ravelli, L.: 1995, A dynamic perspective, *in* Hasan and Fries (1995b), pp. 187–234.

Reboul, A.: 1997, (In)cohérence et anaphore: mythes er réalité, *in* de Mulder et al. (1997), pp. 297–314.

Reinhart, T.: 1981, Pragmatics and linguistics. An analysis of sentence topics, *Philosophica* **27**(1), 53–94.

Resnick, S. I.: 1992, *Adventures in Stochastic Processes*, Birkhäuser, Boston, Basel, Berlin.

Rickheit, G. (ed.): 1991, *Kohärenzprozesse: Modellierung von Sprachverarbeitung in Texten und Diskursen*, Westdeutscher Verlag, Opladen.

Roberts, C.: 1997, Information structure in discourse: Towards an integrated formal theory of pragmatics, *Ohio State University Working Papers in Linguistics* **49**, 91–136.

Rodgers, P. C.: 1966, A discourse-centered rhethoric of the paragraph, *College Composition and Communication* **16**, 2–11.

Rooth, M.: 1996, On the interface principles for intonational focus, *Proceedings of Semantics and Linguistic Theory VI, Rutgers University, New Jersey, April 26-28*.

Ross, K. and Ostendorf, M.: 1996, Prediction of abstract prosodic labels for speech synthesis, *Computer Speech and Language* **10**, 155–185.

Ruhrmann, G.: 1989, *Rezipient und Nachricht: Struktur und Prozesse*, Westdeutscher Verlag, Opladen.

Ruhrmann, G.: 1994, Ereignis, Nachricht und Rezipient, *in* Merten et al. (1994), pp. 237–256.

Russell, B.: 1919/1993, Descriptions, *in* Moore (1993). Excerpt from Chapter XVI of Russell, Bertrand (1919): *Introduction to Mathematical Philosophy*, London: Allen & Unwin.

Sacks, H.: 1995, *Lectures on Conversation*, Blackwell, Oxford. edited by Gail Jefferson.

Sag, I. and Wasow, T.: 1999, *Syntactic Theory: A Formal Introduction*, CSLI Publications, Stanford, CA.

Sanders, T. J., Spooren, W. P. and Noordman, L. G.: 1992, Towards a taxonomy of coherence relations, *Discourse Processes* **15**, 1–35.

Sanford, A. J. and Garrod, S. C.: 1981, *Understanding Written Language. Explorations of Comprehension Beyond the Sentence*, Wiley, London.

Sanford, A. J. and Garrod, S. C.: 1989, What, when, and how?: Questions of immediacy in anaphoric reference resolution, *Language and Cognitive Processes* **4**(3/4), 235–262.

Sanford, A. J. and Garrod, S. C.: 1994, Selective processes in text understanding, *in* Gernsbacher (1994), pp. 699–720.

Sanford, A. J. and Garrod, S. C.: 1998, The role of scenario mapping in text comprehension, *Discourse Processes* **26**(2/3), 159–190.

Sanford, A. J. and Moxey, L. M.: 1995, Aspects of coherence in written language: a psychological perspective, *in* Gernsbacher and Givón (1995), pp. 161–215.

Sankoff, D.: 1978, Probability and linguistic variation, *Synthese* **37**, 217–238.

Sankoff, D. and Labov, W.: 1979, On the uses of variable rules, *Language in Society* **8**, 189–222.

Sasse, H.-J.: 1987, The thetic-categorical distinction revisited, *Linguistics* **25**, 511–580.

Schade, U., Langer, H., Rutz, H. and Sichelschmidt, L.: 1991, Kohärenz als prozeß, *in* Rickheit (1991), pp. 7–58.

Schank, R.: 1977, Rules and topics in conversation, *Cognitive Science* **1**, 421–442.

Schank, R.: 1982, *Dynamic Memory*, Cambridge University Press, Cambridge.

Schenk, M.: 1999, Kommunikationstheorien, *in* Noelle-Neumann et al. (1999), pp. 171–186.

Schiller, M.: 1961, The sheep's in the meadow, *The Antioch Review* **XXI**, 336–340.

Schlobinski, P. and Schütze-Coburn, S.: 1992, On the topic of topic and topic continuity, *Linguistics* **30**, 89–121.

Schmidt, Siegfried, J.: 1994, Die Wirklichkeit des Beobachters, *in* Merten et al. (1994), pp. 3–19.

Schneider, W. and Raue, P.-J.: 1998, *Handbuch des Journalismus*, updated pocket edn, Rowohlt, Reinbek.

Schönbach, K. and Früh, W.: 1984, Der dynamisch-transaktionale Ansatz II: Konsequenzen, *Rundfunk und Fernsehen* **32**, 314–329. reproduced in (Früh 1992b, 41–58).

Schramm, W. and Roberts, D. F. (eds): 1971, *The Process and Effects of Mass Communication.*, University of Illinois Press, Urbana, IL.

Schröder, B.: 1999, *Zur wissenschaftstheoretischen Struktur von formalen Grammatiktheorien*, PhD thesis, Institut für Logik und Grundlagenforschung der Universität Bonn.

Schulz, W.: 1999, Kommunikationsprozeß, *in* Noelle-Neumann et al. (1999), pp. 140–171.

Schütz, A.: 1960, *Vom sinnhaften Aufbau der sozialen Welt*, Julius Springer, Berlin.

Schütze-Coburn, S.: 1994, *Prosody, Syntax and Discourse: Assessing Information Flow in German Conversation*, PhD thesis, University of California, Los Angeles.

Schwarzschild, R.: 1999, GIVENness, Avoid F and other constraints on the placement of focus, *Natural Language Semantics* .

Selting, M.: 1988, The role of intonation on the organization of repair an d problem handling sequences in conversation, *Journal of Pragmatics* **12**, 293–322.

Sgall, P.: 1987, Prague functionalism and topic vs. focus, *in* Dirven and Fried (1987), pp. 169–189.

Sgall, P., Hajičová, E. and Panevová, J.: 1986, *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, D. Reidel, Dordrecht.

Shannon, C. E. and Weaver, W.: 1949, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.

Sidner, C. L.: 1983, Focusing in the comprehension of definite anaphora, *in* M. Brady and R. C. Berwick (eds), *Computational Models of Discourse*, MIT Press, Cambridge, MA, pp. 267–330.

Singer, M.: 1994, Discourse inference processes, *in* Gernsbacher (1994), pp. 479–516.

Soon, W. M., Ng, H. T. and Lim, C. Y.: 1999, Corpus-based learning for noun phrase coreference resolution, *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Md., 21–22 June 1999, pp. 285–291.

Sperber, D. and Wilson, D.: 1995, *Relevance*, Blackwell, Oxford; Cambridge, MA.

Staab, J.: 1990, *Nachrichtenwert-Theorie. Formale Struktur und empirischer Gehalt.*, Alber, Freiburg/München.

Stacton, D.: 1961, *The Judges of the Secret Court*, Pantheon, New York, NY. Excerpt from pages 50–56.

Steedman, M.: 2000a, Information structure and the syntax-phonology interface, *Linguistic Inquiry* **31**(4).

Steedman, M.: 2000b, *The Syntactic Process*, MIT Press / Bradford Books, Cambridge, MA.

Steen, G.: 1999, Genres of discourse and the definition of literature, *Discourse Processes* **28**, 109–120.

Sternefeld, W.: 1993, Anaphoric reference, *in* Jacobs, von Stechow and Sternefeld (1993), pp. 953–963.

Stevenson, R., Crawley, R. and Kleinman, D.: 1994, Thematic roles, focus and the representation of actions, *Language and Cognitive Processes* **9**, 519–548.

Stevenson, R. J.: 1996, Mental models, propositions, and the comprehension of pronouns, *in* Oakhill and Garnham (1996), pp. 53–76.

Straßner, E. (ed.): 1975, *Nachrichten—Entwicklungen, Analysen, Erfahrungen*, Fink, München.

Strube, M.: 1996, *Funktionales Centering*, PhD thesis, Albert-Ludwigs-Universität Freiburg, Freiburg.

Strube, M.: 1998, Never look back: An alternative to centering, *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Québec, Canada, 10–14 August 1998, Vol. 2, pp. 1251–1257.

Strube, M. and Hahn, U.: 1996, Functional centering, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA, 24–27 June 1996, pp. 270–277.

Strube, M. and Hahn, U.: 1999, Functional Centering: Grounding referential coherence in information structure, *Computational Linguistics* **25**(3), 309–344.

Strube, M. and Wolters, M.: 2000, A probabilistic genre-independent model of pronominalization, *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, Wash., 29 April – 3 May 2000, pp. 18–25.

Stümpert, H.: 1991, Radio-Report, *in* Buchholz and LaRoche (1991), pp. 97–101.

Supported Coding Schemes: 1998. MATE Deliverable D1.1.

Suri, L. Z. and McCoy, K. F.: 1994, RAFT/RAPR and centering: A comparison and discussion of problems related to processing complex sentences, *Computational Linguistics* **20**(2), 301–317.

Svartvik, J. (ed.): 1990, *The London-Lund corpus of spoken English*, number 82 in *Lund Studies in English*, Lund University Press.

Swales, J.: 1990, *Genre Analysis*, Cambridge University Press.

Tannen, D.: 1979, *Conversational Style: Analyzing Talk Among Friends*, Ablex, Norwood, NJ.

Tardieu, H., Ehrlich, M.-F. and Gyselinck, V.: 1992, Levels of representation and domain-specific knowledge in comprehension of scientific texts, *Language and Cognitive Processes* **7**(3/4), 335–351.

Teufel, S.: 1998, Meta-discourse markers and problem-structuring in scientific articles, *Proceedings of the COLING-ACL '98 Workshop on Discourse Structure and Discourse Markers, Montréal, August, 1998*.

Teufel, S.: 1999, *Argumentative Zoning: Information Extraction from Scientific Articles*, PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Teufel, S., Carletta, J. and Moens, M.: 1999, An annotation scheme for discourse-level argumentation in research articles, *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 8–12 June 1999, pp. 110–117.

Thorsen, N. G.: 1985, Intonation and text in Standard Danish, *Journal of the Acoustical Society of America* **77**, 1205–1216.

Tomlin, R. S.: 1987a, Linguistic reflections of cognitive events, *in* R. S. Tomlin (ed.), *Coherence and Grounding in Discourse*, John Benjamins, Amsterdam; Philadelphia, PA, pp. 455–480.

Tomlin, R. S. (ed.): 1987b, *Coherence and Grounding in Discourse*, John Benjamins, Amsterdam; Philadelphia, PA.

Toole, J.: 1996, The effect of genre on referential choice, *in* Fretheim and Gundel (1996), pp. 263–290.

Toulmin, S.: 1958, *The Uses of Argument*, Cambridge University Press, Cambridge.

Trabasso, T. and Suh, S.: 1993, Understanding text: Achieving explanatory coherence thorugh on-line inferences and mental operations in working memory, *Discourse Processes* **16**, 3–34.

Traum, D.: 1994, *A Computational Theory of Grounding in Natural Language Conversation*, PhD thesis, University of Rochester.

Trubetzkoy, N. S.: 1939/1989, *Grundzüge der Phonologie*, Vandenhoek & Ruprecht, Göttingen.

Tulving, E.: 1985, How many memory systems are there?, *American Psychologist* **40**, 385–398.

Turan, Ü. D.: 1995, *Null vs. Overt Subjects in Turkish: A Centering Approach*, PhD thesis, University of Pennsylvania, Philadelphia, PA.

Turan, Ü. D.: 1998, Ranking forward-looking centers in Turkish: Universal and language specific properties, *in* Walker et al. (1998), pp. 138–160.

Ungeheuer, G.: 1967/1972a, *Die kybernetischen Grundlagen der Sprachtheorie von Karl Bühler*, in Ungeheuer (1972c), pp. 171–190.

Ungeheuer, G.: 1970/1972b, *Kommunikative und extrakommunikative Betrachtungsweisen in der Phonetik*, in Ungeheuer (1972c), pp. 184–222.

Ungeheuer, G.: 1972c, *Sprache und Kommunikation*, Buske, Hamburg.

Ungeheuer, G.: 1974/1987b, *Kommunikationssemantik: Skizze eines Problemfeldes*, in Ungeheuer (1987c), pp. 70–100. edited by Johann Georg Juchem.

Ungeheuer, G.: 1974/1987e, *Was heißt Verständigung durch Sprechen?*, in Ungeheuer (1987c), pp. 34–99. edited by Johann Georg Juchem.

Ungeheuer, G.: 1980/1987a, *Gesprächsanalyse an literarischen Texten*, in Ungeheuer (1987c), pp. 184–222. edited by Johann Georg Juchem.

Ungeheuer, G.: 1982/1987d, *Vor-Urteile über Sprechen, Mitteilen, Verstehen*, in Ungeheuer (1987c), pp. 290–338. edited by Johann Georg Juchem.

Ungeheuer, G.: 1987c, *Sprechen, Mitteilen, Verstehen*, Rader, Aachen. edited by Johann Georg Juchem.

Vallduvi, E.: 1990, *The Informational Component*, PhD thesis, University of Pennsylvania, Department of Linguistics, Philadelphia, Penn.

Vallduvi, E. and Engdahl, E.: 1996, The linguistic realization of information packaging, *Linguistics* **34**, 459–519.

Vallduví, E. and Vilkuna, M.: 1998, On rheme and kontrast, *in* Culicover and McNally (1998), pp. 79–108.

van den Bosch, A.: 1997, *Learning to pronounce written words: A case study in inductive language learning*, Philippides.

van den Broek, P.: 1994, Comprehension and memory of narrative texts: Inference and coherence, *in* Gernsbacher (1994), pp. 539–588.

van Dijk, T. A.: 1972, *Some Aspects of Text Grammars*, Mouton, The Hague.

van Dijk, T. A.: 1977, *Text and Context. Explorations in the Semantics and Pragmatics of Discourse*, Longman, London.

van Dijk, T. A.: 1980, *Textwissenschaft. Eine interdisziplinäre Einführung*, Deutscher Taschenbuch Verlag, München. Dutch original: Tekstwetenschap. Een interdisciplinaire inleiding. Utrecht/Antwerpen: Het Spectrum 1978.

van Dijk, T. A.: 1981, Sentence topic versus discourse topic, *Studies the Pragmatics of Discourse*, Mouton, The Hague; Berlin, pp. 177–194.

van Dijk, T. A.: 1985a, Structures of news in the press, *in* van Dijk (1985b), pp. 69–93.

van Dijk, T. A. (ed.): 1985b, *Discourse and communication*, de Gruyter, Berlin.

van Dijk, T. A. (ed.): 1985c, *Handbook of Discourse Analysus*, Vol. 2, Dimensions of Discourse, Academic Press, London.

van Dijk, T. A. and Kintsch, W.: 1983, *Strategies of Discourse Comprehension*, Academic Press, London.

van Hoek, K.: 1995, Conceptual reference points: A cognitive grammar account of pronominal anaphora constraints, *Language* **71**(2).

van Kuppevelt, J.: 1996, Directionality in discourse: prominence differences in subordination relations, *Journal of Semantics* **13**(4), 363–395.

van Leeuwen, T.: 1984, Impartial speech: observations on the intonation of newsreaders, *Australian Journal of Cultural Studies* **2**(1), 84–98.

Veenstra, J., van den Bosch, A., Daelemans, W., Buchholz, S. and Zavřel, J.: 2000, Memory-based word-sense disambiguation, *Computing and the Humanities* **34**(1/2).

Venables, W. N. and Ripley, B. D.: 1997, *Modern Applied Statistics with S-PLUS*, 2 edn, Springer, New York, NY.

Vieira, R.: 1998, *Definite Description Resolution in Unrestricted Texts*, PhD thesis, University of Edinburgh.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D. and Hirschman, L.: 1995, A model-theoretic coreference scoring scheme, *Proceedings fo the 6th Message Understanding Conference (MUC-6)*, Morgan Kaufmann, San Mateo, Cal., pp. 45–52.

Vion, M. and Colas, A.: 1999, Expressing coreference in french: Cognitive constraints and development of narrative skills, *Journal of Psycholinguistic Research* **28**(3), 261–291.

von der Gabelentz, G.: 1891, *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse*, T.O. Weigel Nachfolger, Leipzig.

von Heusinger, K.: 1997, *Salienz und Referenz*, Akademie Verlag, Berlin.

von Heusinger, K.: 1999, Intonation and information structure. Habilitationsschrift, Universität Konstanz.

von Stechow, A. and Wunderlich, D. (eds): 1991, *Handbuch Semantik*, Walter de Gruyter, Berlin; New York, NY.

Vonk, W., Hustinx, L. G. and Simons, W. H.: 1992, The use of referential expressions in structuring discourse, *Language and Cognitive Processes* **7**(3/4), 301–333.

Vossen, P. (ed.): 1998, *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer, Dordrecht.

Wachtel, S.: 1997, *Schreiben fürs Hören*, UVK Medien, Konstanz.

Walker, M. A.: 1993, *Informational Redundancy and Resource Bounds in Dialogue*, PhD thesis, University of Pennsylvania, Department of Computer and Information Science, Pennsylvania, Penn.

Walker, M. A.: 1996, Limited attention and discourse structure, *Computational Linguistics* **22**(2), 255–264.

Walker, M. A.: 1998, Centering, anaphora resolution, and discourse structure, *in* Walker et al. (1998), pp. 401–435.

Walker, M. A.: 2000, Toward a model of the interaction of centering with global discourse structure, *Verbum* .

Walker, M. A., Iida, M. and Cote, S.: 1994, Japanese discourse and the process of centering, *Computational Linguistics* **20**(2), 193–233.

Walker, M. A., Joshi, A. K. and Prince, E. (eds): 1998, *Centering Theory in Discourse*, Oxford University Press.

Walker, M. A. and Prince, E.: 1996, A bilateral approach to givenness: A hearer-status algorithm and a centering algorithm, *in* Fretheim and Gundel (1996), pp. 291–306.

Ward, G., Sproat, R. and McKoon, G.: 1991, A pragmatic analysis of so-called anaphoric islands, *Language* **67**(437–474).

Watzlawick, P., Beavin, J. and Jackson, D.: 1967, *Pragmatics of Human Communication*, Norton, New York, NY.

Webber, B. L.: 1981, Discourse model synthesis: preliminaries to reference, *in* Joshi et al. (1981), chapter 13, pp. 283–299.

Webber, B. L.: 1983, So what can we talk about now?, *in* M. Brady and R. C. Berwick (eds), *Computational Models of Discourse*, MIT Press, Cambridge, MA, pp. 331–371.

Webber, B. L.: 1991, Structure and ostension in the interpretation of discourse deixis, *Language and Cognitive Processes* **6**(2), 107–135.

Weber, M.: 1940, *Wirtschaft und Gesellschaft*.

Wegener, P.: 1885, *Grundfragen des Sprachlebens*, Niemeyer, Halle.

Weil, H.: 1844/1978, *The Order of Words in the Ancient Languages compared wiht that of the Modern Languages*, John Benjamins, Amsterdam. Translated by Charles W. Super.

Weinrich, H.: 1976, *Textlinguistik: Zur Syntax des Artikels in der deutschen Sprache*, Klett, Stuttgart, chapter IX, pp. 163–176.

Wettschereck, D., Aha, D. W. and Mohri, T.: 1997, A review and empirical evaluation of feature-weighting methods for a class of lazy learning algorithms, *Artificial Intelligence Review* **11**, 273–314.

White, D. M.: 1950, The gatekeeper: A case study in the selection of news, *Journalism Quarterly* **27**, 383–390.

Wiebe, J. M.: 1991, References in narrative text, *Noûs* **25**(4), 457–486.

Wiebe, J. M.: 1994, Tracking point of view in narrative, *Computational Linguistics* **20**(2), 233–287.

Wiebe, J. M., Bruce, R. F. and O'Hara, T. P.: 1999, Development and use of a gold-standard data set for subjectivity classification, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, Md., 20–26 June 1999, pp. 246–253.

Wiesemann, U. (ed.): 1986, *Pronominal Systems*, Gunter Narr, Tübingen.

Wilke, J. (ed.): 1990, *Fortschritte der Publizistikwissenschaft*, Alber, Freiburg.

Winston, P. H. (ed.): 1975, *The psychology of computer vision*, McGraw-Hill, New York, NY.

Wittich, B. (ed.): 1976/1986, *Zeugen liegen bei*, Deutscher Taschenbuch Verlag, München.

Wolters, M.: 1999, Prosodic correlates of referent status, *Proceedings of the XIV. International Congress of Phonetic Sciences, San Francisco, CA.*

Wolters, M.: in preparation, Psychological subjects from a psychological perspective. Institut für Kommunikationsforschung und Phonetik, Unversität Bonn.

Wolters, M. and Kirsten, M.: 1999, Exploring the use of linguistic features in domain and genre classification, *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, 8–12 June 1999, pp. 142–149.

Wolters, M. and Mixdorff, H.: 2000, Evaluating radio news intonation: Autosegmental versus superpositional modelling, *Proceedings of the International Conference on Spoken Language Processing, Beijing, China*.

Wright, S. and Givón, T.: 1987, The pragmatics of indefinite reference: quantified text-based studies, *Studies in Language* **11**(1).

Wrobel, S.: 1996, Inductive logic programming, *in* Brewka (1996), pp. 153–189.

Young, S. and Bloothooft, G. (eds): 1997, *Corpus-Based Methods in Language and Speech Processing*, Kluwer, Boston.

Zadrozny, W. and Jensen, K.: 1991, Semantics of paragraphs, *Computational Linguistics* **17**(2), 171–209.

Zehrt, W.: 1996, *Hörfunk-Nachrichten*, Ölschläger in UVK, Konstanz.

Zinsser, W. K.: 1997, *On Writing Well*, sixth, revised and updated edn, HarperCollins, New York, NY.

Zwicky, A., Salus, P., Binnick, R. I. and Vanek, A. (eds): 1971, *Studies out in the left field. Defamatory essays presented to Jim McCawley*, Linguistic Research, Edmonton, Canada.

List of Abbreviations

ANOVA	analysis of variance
C/I theory	Construction/Integration theory of discourse comprehension
C_b	backward-looking centre of Centering Theory, c.f. Section 4.3.2
C_f	forward-looking centre of Centering Theory, c.f. Section 4.3.2
DLF	Deutschlandfunk (German radio station, now Deutschlandradio)
DS	Discourse segment
DSP	Discourse segment purpose (in the theory if Grosz and Sidner 1986)
GB	Government and Binding Theory
GLM	generalised linear model
GS	Grosz and Sidner's model of discourse structure
LCS	Lexical Conceptual Structure (Jackendoff 1990)
MCU	Major Clause Unit, definition c.f. page 111
MOP	Memory Organization Packet (Schank 1982)
MUCCS	Message Understanding Conference Co-Reference Scheme
NP	Noun Phrase
PET	Personal Experience Theory, c.f. Appendix D
RD	Referential Distance (c.f. p. 85
RST	Rhetorical Structure Theory (c.f. Section 3.3.2)
SGML	Standard Generalised Markup Language (Goldfarb 1990)
SFL	Systemic Functional Linguistics (Halliday 1994)
SMF	Scenario Mapping and Focus Theory (Sanford and Garrod 1981, Sanford and Garrod 1998)
TOP	Thematic Organization Packet (Schank 1982)
TP	Topic Persistence (c.f. page 85
VP	Verb Phrase
WBUR	Boston public radio station

Appendix A Analysed Texts

This appendix collects two longer texts which are analysed in the main bodu of the thesis: a newspaper story about the assassination of Bashir Gemayel (Section A.1), and the opening paragraphs of Terry Pratchett's novel "Guards, Guards" (taken from (Pratchett 1990), Section A.2).

A.1 The Gemayel Text

In the following text, footnotes mark the end of van Dijk's macrostructural boundaries. All referring expressions are marked by square brackets; paragraphs are numbered in round brackets. The letter codes, summarised in Table A.1 mark members of co-specification sequences.

(1) [Israeli troops]$_{IT}$ re-enter [west Beirut]$_{WB}$

HEADLINE

(2) *BEIRUT - [Israeli forces]$_{IT}$ moved into [west Beirut]$_{WB}$ yesterday to "insure quiet" after [the assassination of [Lebanese president-elect Bashir Gemayel]$_G$]$_{GA}$, [the Israeli military command in [Jerusalem]$_J$]$_{IM}$ said.*LEAD

(3) [Unidentified assassins] killed [Gemayel]$_G$ [Tuesday] with [a 204-kg (450 lb) bomb that took [more than 26 lives], wounded [60 other people] and returned [Lebanon]$_L$ to [relentless sectarian violence]].MAIN EVENT

(4) "As a result of [the assassination of [Bashir Gemayel]$_G$]$_{GA}$, [Israel Defence Forces]$_{IT}$ entered [west Beirut]$_{WB}$ in order to prevent [possible severe occurrences] and in order to insure

People and Events			
G	Gemayel	A	Yassir Arafat
GA	Gemayel's assassination	B	Menachem Begin
GD	Gemayel's death	IT	Israeli troops
GB	blast in assassination	IM	Israeli military command
GBo	Gemayel's body	W	Prime Minister Wazzan
GN	news of G.'s assassination	D	Envoy Draper
Places			
I	Israel	J	Jerusalem
L	Lebanon	WB	West Beirut

Table A.1. Codes for important discourse entities in Gemayel text

[quiet]," [a statement by [the Israeli military command]$_{IM}$] said.

CONSEQUENT ACTION 1

(5) [The death of [the Maronite Christian]$_G$]$_{GD}$, only nine days before [he]$_G$ was to be inaugurated as [Lebanon's]$_L$ president, raised [fears of [a new round of fighting between [[Gemayel's]$_G$ troops] and [Muslim forces] in [the deeply divided country]$_L$]].

EXPECTATIONS

(6) [The Government], shocked at [the first assassination in [Lebanese history] of [a person elected president]$_G$]$_{GA}$, delayed confirming [the death of [the 34-year-old right-wing leader]$_G$]$_{GD}$ for [nine hours].

MAIN EVENT (CONTINUED)

(7) [All crossings between [east and west Beirut]] were closed and [panicky residents] jammed [gas stations] and [bakeries] stocking up in fear [a continued closure] would lead to [shortages of [essential items]].

CONSEQUENT EVENTS

(8) [An Israeli Army spokesman] said [the border between [Israel]$_I$ and [Lebanon]$_L$ was sealed off yesterday for all but [military personnel], barring [journalists and other civilians] from crossing [the frontier].

(9) "With [great pain] [I]$_W$ face [this shocking news]$_{GN}$ with [the strongest denunciation for [this criminal act]$_{GA}$]," [Prime Minister Chefik Wazzan]$_W$ said [late Tuesday] in [an official statement about [[Gemayel's]$_D$ death]$_{GD}$].

VERBAL REACTION

(10) [President Elias Sarkis] ordered [seven days of [official mourning]] and [a state funeral yesterday in [[Gemayel's]$_G$ hometown of Bikfaya]].

(11) [Six hours after [the blast]]$_{GB}$, [[Gemayel's]$_G$ mangled body]$_{GBo}$ was pulled from [the rubble]. [Government sources] said [it]$_{GBo}$ could only be identified by [[his]$_G$ ring].

MAIN EVENT (CONTINUED)

PLOT

(12) Despite [the charges of [a plot]], no one claimed responsibility for [the blast]$_{GB}$.

(13) [Gemayel]$_G$ was elected over [the protests of [most Muslims, who remembered [[his]$_G$ role as the Phalangist military commander during [the bitter 1975-76 civil war]]].

HISTORY

(14) Twice before – in [March 1979] and [February 1980] – [enemies] tried to kill [Gemayel]$_G$ with [car bombs]. [The second blast] killed [[his]$_G$ 18-month-old daughter].

HISTORY

(15) "[The news of [the cowardly assassination]$_{GA}$]$_{GN}$ . . . is a shock to [the American people and to civilised men and women] everywhere," [President Reagan] said in [a statement from [the White House]].

VERBAL REACTION 1

CRIMINAL

(16) In [Jerusalem]$_J$, [Israeli Prime Minister Menachem Begin]$_B$ cabled [[his]$_B$ condolences] to [[Gemayel's]$_G$ father, Pierre], saying [he]$_B$ was "shocked to [the depths of [[my]$_B$ soul] at [the criminal assassination]$_{GA}$."

VERBAL REACTION 2

(17) [US mideast envoy Morris Draper]$_D$ yesterday met with [Begin]$_B$ in [Jerusalem]$_J$ and vowed to negotiate an [Israeli and Syrian withdrawal] from [Lebanon]$_L$ despite [complications

caused by [[Gemayel's]$_G$ death]$_{GD}$].
CONTEXT

(18) [[Begin's]$_B$ Press spokesman, Uri Porath], said [[Begin]$_B$ and [Draper]$_D$] agreed to work out [a timetable for [the withdrawal of [all foreign forces from [Lebanon]$_L$]]].

(19) Meanwhile in [Rome], [PLO chairman Yasser Arafat]$_A$ yesterday urged [Israel]$_I$ to "return to [[its]$_I$ senses]" and negotiate for [a peaceful settlement of [the Middle East conflict]].

(20) In [a 19-minute speech at [the Inter-Parliamentary Union]], boycotted by [Israeli delegates], [Arafat]$_A$ blamed [Israel]$_I$ for [the murder of [Gemayel]$_G$]$_{GA}$ and called on [the parliamentarians] to set up [a special panel to investigate [[Israel's]$_I$ "war crimes" in [Lebanon]$_L$]. [He]$_A$ accused [Israel]$_I$ of trying to turn [Lebanon]$_L$ into "a protectorate" – UPI, AP[1]

A.2 Guards, Guards

Section 1:

1. This is where the dragons went.

2. They lie . . .

3. Not dead, not asleep. Not waiting, because waiting implies expectation. Possibly the word we're looking for here is . . .

4. . . . dormant. And although the space they occupy isn't like normal space, nevertheless they are packed in tightly. Not a cubic inch there but is filled by a claw, a talon, a scale, the tip of a tail, so the effect is like one of those trick drawings and your eyeballs eventually realise that the space between each dragon is, in fact, another dragon.

5. They could put you in mind of a can of sardines, if you thought sardines were huge and scaly and proud and arrogant.

6. And presumably, somewhere, there's the key.

Section 2:

1. In another space entirely, it was early morning in Ankh–Morpork, oldest and greatest and grubbiest of cities. A thin drizzle dripped from the grey sky and punctuated the river mist that coiled among the streets. Rats of various species went about their nocturnal occasions. Under night's damp cloak assassins assassinated, thieves thieved, hussies hustled. And so on.

2. And drunken captain Vimes of the Night Watch staggered slowly down the street, folded gently into the gutter outside the Watch House and lay there while, above him, strange letters made of light sizzled in the damp and changed colour . . .

[1]The news agencies that supplied the information

3. The city wasa, wasa, wasa wossname. Thing. *Woman*. Thass what it was. Woman. Roaring, ancient, centuries old. Strung you along, let you fall in thingy, love, with her, then kicked you inna, inna, tingy. Thingy, in your mouth. Tongue. Tonsils. *Teeth*. That's what it, she, did. She wasa . . . thing, you know, lady dog. Puppy. Hen. *Bitch*. And then you hated her and, and just when you thought you'd got her, it, out of your, your, whatever, then she opened her great booming rotten heart to you, caught you off bal, bal, bal, thing. *Ance*. Yeah. Thassit. Never know where where you stood. Lay. Only thing you were sure of, you couldn't let her go. Because, because she was yours, all you had, even in her gutters . . .

Section 3:

1. Damp darkness shrouded the venerable buildings of Unseen University, premier college of wizardry. The only light was a faint octarine flicker from the tiny windows of the new High Energy Magic building, where keen-edged minds were probing the very fabric of the universe, whether it liked it or not.

2. And there was light, of course, in the Library.

3. The Library was the greatest assemblage of magical texts anywhere in the multiverse. Thousands of volumes of occult lore weighted its shelves.

4. It was said that, since vast amounts of magic can seriously distort the mundane world, the Library did not obey the normal rules of space and time. It was said that it went on *forever*. It was said that you could wander for days among the distant shelves, that there were lost tribes of research students somewhere in there, that strange things lurked in forgotten alcoves and were preyed on by other things that were even stranger.[2]

5. Wise students in search of more distant volumes took care to leave chalk marks on the shelves as they roamed deeper into the fusty darkness, and told friends to come looking for them if they weren't back by supper.

6. And, because magic can only loosely be bound, the Library books themselves were more than merepulped wood and paper.

7. Raw magic crackled from their spines, earthing itself harmlessly in the copper rails nailed to every shelf for that very purpose. Faint traceries of blue fire crawled across the bookcases and there was a sound, a papery whispering, such as might come from a colony of roosting starlings. In the silence of the night the books talked to one another.

8. There was also the sound of someone snoring.

[2]All this was untrue. The truth is that even big collections of ordinary books distort space, as can readily be proved by anyone who has been around a really old-fashioned secondhand book shop, one of those that look as though they were designed by M. Escher on a bad day and has more staircases than storeys and those rows of shelves which end in little doors that are surely too small for a full-sized human to enter. The relevant equation is: Knowledge = power = energy = matter = mass; a good bookshop is just a genteel Black Hole that knows how to read.

9. The light from the shelves didn't so much illuminate as highlight the darkness, but by its violet flicker a watcher might just have identified an ancient and battered desk right under the central dome.

10. The snoring was coming from underneath it, where a piece of tattered blanket barely covered what looked like a heap of sandbags but was in fact an adult male orangutan.

11. It was the Librarian.

12. Not many people these days remarked upon the fact that he was an ape. The change had been brought about by magical accident, always a possibility where so many powerful books are kept together, and he was considered to have got off lightly. After all, he was still basically the same shape. And he had been allowed to keep his job, which he was rather good at, although 'allowed' is not really the right word. It was the way he could roll his upper lip back to reveal more incredibly yellow teeth than any other mouth the University Council had ever seen before that somehow made sure the matter was never really raised.

13. But now there was another sound, the alien sound of a door creaking open. Footsteps padded across the floor and disappeared amongst the clustering shelves. The books rustled indignantly, and some of the larger grimoires rattled their chains.

14. The Librarian slept on, lulled by the whispering of the rain.

15. In the embrace of his gutter, half a mile away, Captain Vimes of the Night Watch opened his mouth and started to sing.

Appendix B Statistical Background

Many of the statistical tools I used throughout this thesis will not be familiar to the average linguist even if she specialises in computational or corpus linguistics. Although they are standard fare in other fields, such as sociology, biology or medicine, (and quite old hats in statistics, to be honest) they are only encountered here and there in corpus studies. Unsurprisingly, the most accessible and thorough introduction to these methods I have found so far is (Lindsay 1995), which was written for social science students. Therefore I have decided to discuss these methods in somewhat more detail than usual. I assume little previous knowledge; readers should merely know what a mean, a variance, and a probability distribution is.

This appendix is structured as follows: First I critically discuss the role of statistical analysis in corpus research (Section B.1). Then I introduce the concept of random variables (Section B.2). On this basis I then explain how generalised linear models can be used to describe the distribution of a random variable (Section B.3). Section B.4 focuses on two measures of association, $\lambda_{\max}$ and Goodman and Kruskal's τ. Finally, in section B.5, I introduce two basic types of stochastic processes, Poisson processes and Markov Chains, which are needed in Chapter 5.4.

B.1 Statistical Analysis of Corpora

In this section I address three general issues that needed to be addressed for the corpus analyses reported in this thesis: the choice of statistical tests, the reliability of the analysed data, and the limits of purely corpus-based results.

Choice of Tests: Parametric or Non-Parametric? Most of the well-established statistical methods such as the t-test are *parametric*. These methods have been developed for interval-scaled data that show a normal distribution - which is *not* the case for language data. First and most important of all, language data is *categorical*. In most cases we cannot establish a rank order between instances of a linguistic variable. Take for example the variable "forms of referring NPs". It is nonsense to say that definite NPs are more of a referring NP than pronouns. On the other hand a variable that *can* be regarded as ordinal is the accessibility of a discourse entity, because we can say that an entity e_1 that has not been mentioned for the last 2 paragraphs is less accessible than an entity e_2 that was last mentioned a sentence ago. However, accessibility is still not interval-scaled because it would be nonsense to say that e_1 is four or five times less accessible than e_2. Even when we can specify continuous distributions that approximate discrete linguistic data, these distributions are rarely Gaussian. As a consequence, I will only use non-parametric tests. Most approaches I use are designed for

categorical variables; ordinal variables almost never occurred.

Measuring Annotation Quality: Annotations that are to be analysed statistically need to be reliable and valid (Krippendorff 1980). *Validity* means that annotators are consistent with themselves: when re-annotating a text after a while, they should make the same decisions as in the first annotations. This shows that they have developed stable internal categories, and that the researcher has been able to define the categories in such a way that they do not depend too much on the annotators' personal experience theory (c.f. Section 2.2.2).

An annotation scheme is *reliable* if two or more annotators working according to the same scheme produce annotations which do not diverge greatly from each other (Krippendorff 1980). A common measure of divergence is κ (Cohen 1960, Carletta 1996), which is defined as

$$\kappa = \frac{p(A) - p(E)}{1 - p(E)} \tag{B.1}$$

where p(A) is the probability that the annotators agree on their annotations, and p(E) the chance level of agreement. If the annotators agree perfectly, κ is 1, if agreement is exactly at chance level (p(A) = p(E)), it is 0. A κ below 0.67 signals bad inter-annotator agreement, and a κ above 0.8 indicates that a satisfactory annotation scheme has been found. In this thesis, κ was used for determining the reliability of the Sortal Class annotations documented in Appendix C.1.

Only Corpus Analysis? Statistics is the key to any quantitative corpus analysis. A survey of corpus-based work in recent ICAME, ACL and COLING proceedings shows three groups of papers: those that merely list the contexts in which the items of interest occurred or reporting contingency tables, those that test the significance of certain associations with very standard parametric tests, such as z-scores, t-scores, and χ^2-tests, and those that apply mathematically more complex models, such as factor analysis (Biber 1988) or log-linear models (de Haan 1987). For some scholars, their preferred method of statistical analysis has even become some kind of signature. Which of these groups uses the best methods? The answer depends, as always, on the research questions. In order to make that sibyllinic answer more precise, let us consider the two main problems of statistical corpus analysis: sparse data and limited control.

By *limited control*, I mean that we cannot be sure that the forms we are interested in occur in all relevant contexts of usage. This means that we cannot control for all potential influences as thoroughly as we would in a normal experiment. Moreover, the corpus design criteria will inevitably affect the usage contexts we *do* see and, as a consequence, the distribution of the forms we are examining. These problems cannot be resolved if we just gather instances from the corpus in a principled way, resembling a controlled experiment. The corpus, the source of our data, is intrinsically biased; there are no fully representative corpora, and there will never be (Bergenholtz and Mugdan 1989), since it is impossible to get a balanced view of how and in which contexts language is used in any one moment in time. The only way to deal with this bias in a statistically sound manner is to take it into account in our interpretation.

By *sparse data* I mean that when examining rare linguistic forms, such as stressed pronouns, chances are that the corpus will contain only a few instances of these forms. The problem is exacerbated when it comes to very specific constructions that are very rare in natural speech, but are nevertheless key tests of semantic and syntactic theories, such as second-occurrence

focus (Rooth 1996, Beaver, Clark and Wolters in preparation). We may find too few instances for statistical tests which demand a certain number of instances per conditions in order to be applicable. The necessary size of the corpus crucially depends on the analysis task. But little real progress has been made on this issue except for speculations such as the more frequent the form under study, the smaller the corpus can be (de Haan 1992). The only remedy is to be extremely careful with any inferential statistics. For the corpus analyses in this thesis, I relied heavily on log-linear models and logistic regression (Lindsay 1995, Agresti 1990, Andersen 1990). Both methods, which are based on the theory of generalised linear models, allow to formulate testable hypotheses about the interaction between linguistic variables in corpora.

But not only the statistical interpretation of rare usages is difficult. Rare forms may also be rare because they constitute performance errors, or because they are difficult to process or produce. The time-honoured heuristic "if it occurs frequently enough, it will be acceptable (somehow)" is not applicable anymore. This was one of the reasons why Greenbaum and Quirk supplemented the Survey of English Usage, which was designed to form the basis of a comprehensive grammar of English, (Quirk, Greenbaum, Leech and Svartvik 1985), with elaborate sets of elicitation and judgement experiments (Greenbaum and Quirk 1970).

Thus, if we find a referring expression in an atypical context, we first need to judge whether that referring expression is not in fact misused. But what exactly is misuse? Overly ambiguous co-specifications? Or cases where an anaphoric expression can only be resolved wrongly, even if taking the semantics of the sentence into account? The statistical model covers such errors under the heading "random variation", and a detailed analysis then has to ascertain how much of that variation is due to error. This detailed analysis consists of two steps: inspecting the data that causes the variation in the corpus, and designing experiments to test which of the unusual variants are not acceptable, and why. Such experiments crop up more and more frequently in the literature, be it to replace linguists' intuitions about examples by judgements from untrained native speakers (Bard, Robertson and Sorace 1996, Cowart 1997, Keller 1998), be it to supplement corpus results (de Moennink 1997). For the purposes of this thesis, I limited myself to laying some conceptual foundations and testing these foundations in corpus studies. Designing a test suite for pronoun uses is a complex enough project in itself, and clearly beyond the scope of this thesis.

B.2 Random Variables

Random variables are variables X that take on each of their values v_i with a certain probability $P(X = v_i)$. For example, let R be a random variable whose values describe the form of a referring expression — whether it is headed by a definite (DEF), an indefinite (INDEF), or a demonstrative article (DEM), whether it is a pronoun (PRO), or whether it is a bare NP (BARE). Now, imagine we draw an referring expression at random from paragraphs (12)–(15) of the Gemayel text (Appendix 6.4.3). Those sentences contain 18 referring expressions, of which 9 are definites, 4 proper names, and 2 pronouns. We can estimate the probability that this referring expression will be definite (R = DEF) by the percentage of definite descriptions in the text. This is 50.0%, so that the empirical estimate for P(R = DEF), $\hat{P}$(R = DEF), is 0.5; the circumflex denotes that $\hat{P}$(R = DEF) is an estimate. But is 0.5 this the real, "true" probability that a referring expression drawn randomly from an arbitrary text will be definite? No, for two reasons: First one small

text from one genre does not cover all the linguistic and extralinguistic influences on the form of a referring expression that we would need to take into account. Second, remember that R is a *random* variable: although it is extremely unlikely that we will find no definite referring expressions in an English text when P(R = DEF) = 0.5, this does not mean that this case cannot occur. Mathematically, the P(X = v_i) are determined by a so-called *probability mass function* p(x). This function allows us to make predictions about the relative frequencies of the values v_i.

Most of the random variables we are dealing with here are categorical; what we are modelling quantitatively here are not the categories themselves, but their counts, or, more generally, how often they are likely to occur in arbitrary data sets. It specifies a probability distribution on the values v_i of R. Lindsay (1995, Chapter 4) surveys the most important distributions for describing counts.

In Chapters 6 and 7, the random variables we are interested in describe the form of a referring expression. Most of the aspects we are interested in can be expressed by binary random variables: Pronominalised or not? Definite NP or not? Modifier or not?

Binary random variables are commonly modelled using a binomial probability distribution. Let's assume we have a fixed number of events which occur independently of each other. In our case, the events are referring expressions. We want to determine how many of the events belong to the category we are interested in, say, pronouns; these events count as hits, all others as misses. If the probability of scoring a hit, i.e. finding a pronoun in a text, is p, then the probability of scoring x hits in a sample of n events is

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{B.2}$$

This equation specifies the binomial distribution. In Chapter 7, this probability occurs again, this time with p and n as additional parameters after the x. P(x;n,p) thus is the probability that we will score x hits in n trys, given that the probability of a hit is p.

B.3 Generalised Linear Models for Categorical Data

B.3.1 What is a Generalised Linear Model?

We have just seen the most basic case of a statistical model, where we only want to find out how the random variable we have measured is distributed. Now, we want to know whether the distribution of a variable of interest, the *target* or *response* variable Y, can be explained by a set of *explanatory* or *predictor* variables X_i. In other words, we want a function that predicts the value of Y on the basis of the values of the X_i.

In standard linear regression modelling, where all our random variables have a Gaussian distribution, we assume that this function is linear:

$$y = \sum_i \beta_i x_i \tag{B.3}$$

The lowercase letters correspond to observations of the random variables Y, X_i. The β_i are the parameters of the model. Each term has one parameter. A model is usually specified by

equations of the following form:

(B.4) $$Y \sim X_1 + \ldots + X_n$$

where the X_i are the predictor variables, and Y the target variable. We can also combine the predictor variables into more complex terms on the right hand side of Eq. B.4. A term of the type $X_i \cdot X_j$ stands for a compound variable. The values of this compound variable are all possible combinations of the values of X_i and X_j. X_i:X_j, on the other hand, stands for X_i conditioned on X_j.

If one of the X_i is categorical, the variable needs to be recoded. For our purposes, we recode variables in the standard way: Each categorical variable X with n values $\mathrm{v}_1 \ldots \mathrm{v}_n$ is replaced by $n-1$ binary subvariables $X'_1 \ldots X'_{n-1}$. Variable X'_i is 1 if $X = \mathrm{v}_{n+1}$. If $X = \mathrm{v}_1$, then $X'_1 = X'_2 = \ldots X'_{n-1} = 0$. The parameters β_i are estimated by fitting Eq. (B.3) to the data set. Other recoding conventions are described in e.g. (Venables and Ripley 1997, Bortz 1993, Andersen 1990, Agresti 1990).

Generalised linear models (GLMs) extend linear models to cases where the distribution of the response variable comes from the exponential family, dropping the strict requirement that the distribution be Gaussian. In GLMs, a slightly different function is fitted to the data (McCullagh and Nelder 1983):

(B.5) $$\eta = \sum_i \beta_i x_i; \eta = g(y)$$

The vector y of observations of Y is not fitted directly anymore. Instead, we fit a variable ν, which is derived from the mean μ of y by the *link function* $\nu = g(\mu)$. The form of the link function depends on the probability distribution of y.

In this thesis, the response variable is usually binary. In that case, we can model the response variable as having a binomial distribution and use logistic regression to estimate the coefficients of the model, given a suitable link function. When the response variable has more than two possible values, we will base our model on a Poisson distribution and switch from logistic to log-linear regression. The name of the two types of regression comes from the link function: in the first case it is the logit link, that is the logarithm of x/(1-x), in the second case the natural logarithm. In log-linear modelling, the parameters are estimated on a slightly different background: the aim is no longer to model one target variable by a set of predictor variables, but to estimate the joint distribution of a set of variables. This reformulation of the problem makes sense when it becomes difficult to separate the variables into targets and predictors. In both approaches, regression modelling and the estimation of joint distributions, parameters for the same terms will have the same values.

B.3.2 Model Selection

Every statistical model has to be evaluated against the data it is supposed to describe. The models used here specify probability distributions. Therefore we need ways to compare the predicted distribution against that which we found in the data. There are two options:

1. We generate artificial data from the model distribution and determine whether it deviates significantly from the real data. The two data sets, real and artificial, can be compared by

the χ^2-test or the Kolmogorov-Smirnov test. Given two data sets, these tests compute the probability that they are both drawn from the same population.

2. We measure the distance between our model and the data by calculating the deviance, the deviation between the predicted and the actual values of the contingency table.

On the following pages I will focus on the deviance, a standard measure in model selection. In order to understand how it is calculated, let us recall that when we estimate our statistical model, we are basically fitting it to a contingency tables of counts. These counts specify how often each combination of the values of all variables occur in the data. The saturated model exactly describes the data set: The expected frequencies are the frequencies that actually occur in the cells of the table; there are as many parameters as there are cells in the table. This results in a model that does not generalise, but overfits the data.[1] We now want to determine whether a more parsimonious model is as likely, given the data set, as the saturated model, which of course gives a perfect fit. The likelihood function measures how plausible it is that a model M has generated the observed data set. For the binomial distribution, the likelihood function is

$$L(p;n) = \binom{n}{x} p^x (1-p)^x \tag{B.6}$$

where x is the number of successes, n the number of events, and p the probability that an event is a success. In this case, the model is characterised by the two parameters p and n.

We can now measure the "distance" between two models in terms of the difference of their likelihoods. The most commonly used version of this difference is the *deviance* D:

$$D(M_1, M_2) = -2(L(M_1) - L(M_2)) \tag{B.7}$$

where M_1 is the more general, M_2 the more specific model. For count data this equation becomes

$$D = -2\log \sum_{c \in \text{cells}} n_c \log \frac{p_{c,test}}{p_{c,base}} \tag{B.8}$$

where n_c is the absolute frequency of the combination of variable values that corresponds to cell c in the contingency table, $p_{c,test}$ the relative frequency of that combination as predicted by the Test Model and $p_{c,base}$ the real relative frequency of the combination in the data set. The lower the deviance the better. The deviance is always positive, and can never fall below 0. For more on the deviance, see (Lindsay 1995, Section 3.2).

But the deviance alone cannot be a sufficient criterium for the quality of a statistical model. It can easily be reduced to 0 by applying the saturated model, which does not tell us more than what we knew before. But the saturated model is also the model with the largest number of parameters: it has as many parameters as cells in the contingency table. Therefore Akaike (1974) modified this selection criterion by a term that punishes the number of parameters in the model, yielding the AIC (short for An Information Criterion):

$$AIC = D + 2\#\text{parameters} \tag{B.9}$$

[1] For more on the connection between regression on the predictor variables and fitting the joint distribution of predictor and target variables, see e.g. Andersen (1990, Chapter 8) or Lindsay (1995, Section 2.3.2).

A model M_1 is better than a model M_2 iff its AIC is lower. There is no significance test for the AIC, and Burnham and Anderson (1998) argue that such a test would not make sense, since all the necessary information for comparing the quality of the models is coded in the AIC itself. For more on interpreting the AIC, see in particular (Burnham and Anderson 1998, Chapters 2 and 4).

The deviance is also the basis for the popular likelihood ratio test. For our categorical models, the test statistic of the likelihood ratio test, often called G^2 in the literature, is equivalent to the deviance (Agresti 1990, page 83). The larger G^2, the more the two models to be compared differ. Whether the difference is significant can be estimated using the χ^2 distribution with k degrees of freedom, where k is the number of terms from model M_2 that are omitted in model M_1. For a justification of this approximation, see Andersen (1990).

The deviance is also useful when we want to determine how each of the predictor variables contributes to describing the distribution of the target variable. For linear models, the corresponding technique is the infamous ANOVA, which measures the amount of variation explained by each term and each combination of terms. In ANOVA the variation is specified by the variance. For GLMs, we need to replace the variance by the deviance.

One of the main advantages of regression analysis is that it allows us to evaluate the contribution of each term on the right-hand side of the formula separately. We simply build the model step by step. At each step, a new term from the right-hand side is added to the model. and the deviance of the resulting model is computed. If the deviance of the new, more complex model is not significantly lower than that of the old model, then the new term is irrelevant. Whether the contribution of a term is significant can also be tested using the F-ratio (McCullagh and Nelder 1983, page 69). The F-ratio is given by

$$F = \frac{D(M_1) - D(M_2)}{(df_{M_1} - df_{M_2}) * \sigma^2} \tag{B.10}$$

where M_1 is the more general model with fewer parameters than M_2, M_2 only differs from M_1 in that it includes the term to be tested, σ^2 is the estimated variance, and df_M are the degrees of freedom of the model M. The degree of freedom corresponds to the number of terms that are removed with respect to the saturated model. The significance of the F-ratio is estimated by the F-distribution.

We have seen that categorical data analysis provides powerful tools for formulating hypotheses about data. By comparing different models and performing analyses of deviance, we can determine whether terms in the model are necessary for describing the co-occurrence counts that we find in our data. In order to apply this methodology to our linguistic data, we just need to translate our linguistic factors to random variables and state our hypotheses about the interactions between those factors in terms of a statistical model. But despite these advantages, statistical models are no Swiss army knife for corpus linguistics, because they make strong assumptions about the independence of the events that they model. If the ratio between the mean and the variance of the counts is larger or smaller than 1, this is a sign for *over-dispersion* ($>$ 1) or *under-dispersion* ($<$ 1), and any significance results should be handled with extreme care, because the model is definitely too simplistic.

A Note on the Literature: Although log-linear models and logistic regression are popular tools for the analysis of contingency tables of counts, they are encountered rarely in computational linguistics. In sociolinguistics, log-linear models were introduced as early as 1970 by the seminal work of Sankoff and associates. In that discipline, they go under the name VARBRUL (Sankoff 1978, Kay and McDaniel 1979, Sankoff and Labov 1979). Several linguists have already applied the freely available program to their corpus data, e.g. (Myhill 1992) or (Prince 1992), both on forms of referring expressions. In his survey of statistical methods in corpus linguistics Oakes (1998, Section 1.5) discusses both log-linear modelling and logistic regression briefly. In computational linguistics logistic regression is used for predicting binary variables; recent examples are (Pan 1998), (Franz 1997), and (Kessler et al. 1997). It tends to fare less well than more sophisticated approaches; this might largely be due to the strong independence assumptions behind the binomial distribution. Recently, Bruce and Wiebe (1999) have started to popularise decomposable log-linear models by a review article in Computational Linguistics. They use the program CoCo (Badsberg 1995) for estimating model parameters and comparing different model architectures. Andersen (1990, Chapter 6) demonstrates how hierarchically nested decomposable models can be estimated and analysed with standard software.

B.4 Measures of Association

Most of the time, I use fairly standard tests for detecting associations between two variables: Fisher's Exact Test for contingency tables, or, if the table is too large for my particular combination of program (R, Ihaka and Gentleman 1996) and computer (Pentium 133, 64 MB RAM) and each cell contains enough entries, the χ^2 test. (For 2×2 tables, R's `fisher.test` routine uses the odds ratio test). But in chapter 6, where I deal with small data sets, I supplement these statistical tests with two descriptive measures: Goodman and Kruskal's τ (Agresti 1990, Eq. B.14) and $\lambda_{\max}$(Darlington n.d., Eq. B.15)). Since both measures are little known, here is some background to help understand the numbers.

Goodman and Kruskal's Tau. τ (Eq. B.14) measures how much the amount of unexplained variation of a variable Y, $V(Y)$, is reduced once we know the value of another variable X:

$$\tau = \frac{V(Y) - E[V(Y|X)]}{V(Y)} \tag{B.11}$$

There are several measures for the variation of a categorical variable; for τ, the *Gini concentration* is used:

$$V(X) = \sum_i P(X=i)(1-P(X=i)) = 1 - \sum_i P(X=i)^2 \tag{B.12}$$

where i stands for the possible values of X.

If the variation of Y given X is equal to the variation of Y, there is no association between the two features, the values of X cannot be used to predict Y, hence the proportional reduction is equal to 0. On the other hand, if Y can be predicted completely on the basis of X, then

E[V($Y|X$)] = 0 and the index becomes 1. With

$$E[V(Y|X)] = 1 - \sum_i \sum_j \frac{P(X=i, Y=j)^2}{P(X=i)} \tag{B.13}$$

we then get the following formula for τ:

$$\tau = \frac{\sum_j \sum_i \frac{P(X=i,Y=j)^2}{P(X=i)} - \sum_j P(Y=j)^2}{1 - \sum_j P(Y=j)^2} \tag{B.14}$$

τ has an intuitive interpretation if we predict the value of Y using a *proportional prediction rule*. Proportional prediction means that a new data item is assigned to category i with probability P($Y = i$). In this case, the variation V(Y) gives the probability of an incorrect guess, and τ measures to what degree knowing the value of X reduces the probability of guessing wrong. We will use τ to determine whether it is possible to predict entity status from certain linguistic features in isolation.

Lambda max. λ_{max} has two interpretations, a causation and a prediction one. Here, we will use it as an indicator of causation. Let us assume that X is our suspected "cause" or, more neutrally, the treatment variable, and Y the effect or dependent variable. Assume that the values X correspond to the columns, and the values of Y to the rows of the contingency table.

To compute λ_{max}, we first determine the highest $P(Y = j|X = i)$ for each j. The value i of X for which $P(Y = j|X = i)$ is highest can be said to have the strongest causal links to the value j of Y. In other words, if we want to predict when Y takes the value j on the basis of the current value of X, then $X = i$ is the most reliable indicator of $Y = j$ that we have. If the probability is 1, then the prediction is perfect, if it is lower, there is still some degree of insecurity.

We then pool the $P(Y = j|X = i)$ "best indicators" by summing them up. 1 is subtracted from that sum, and the result is divided by the number of columns n_X minus 1. The formula thus reads:

$$\lambda_{max} = \frac{\sum_j max_i P(Y=j|X=i) - 1}{n_X - 1} \tag{B.15}$$

λ_{max} is 0 if and only if X and Y are independent, in which case we have P($Y|X$)=P(Y), and the sum in the denominator becomes equal to 1. The maximum value λ_{max} can attain is n_Y-1/n_X-1, where n_Y is the number of possible values of Y, and n_X the number of possible values of X. It follows that we cannot compare the λ_{max} values directly for features with different numbers of values. Therefore I never give just the value of λ_{max}, but also the maximum value it can attain and the percentage of the maximum value attained. In computing λ_{max}, the contributions of each category i of X are not weighted according to their frequencies P($X = i$).[2]

[2] λ_{max} also has a frequency interpretation: It compares the observed frequency of a pair of values to the frequency they would have if X and Y were independent and expresses the difference between these two frequencies in terms of the maximally attainable difference, which we get if each value of X is mapped deterministically onto a value of Y.

B.5 Stochastic Processes

This section introduces two types of stochastic processes that are be needed in Section 5.4.1. First I introduce with Poisson processes, then, I briefly present Markov Chains.

B.5.1 Poisson Processes

Poisson processes belong to the family of *point processes*, which describe the distribution of points in space or events in time. Poisson-based models assume that these events are generated by a random mechanism. The times or temporal distances between two events i and $i+1$ are modelled by random variables X_i. These X_i do not depend on each other: They are independently distributed. Furthermore, we assume that the random distribution which determines the length of each X_i is the same for all temporal distances: The X_i are identically distributed. That distances are identically and independently distributed (or i.i.d., for short) is a very strong claim, and it is the first assumption that has to be verified before we can claim that something can be modelled by a Poisson process.

The assumption of i.i.d. distances is something that Poisson processes share with other types of stochastic processes, such as renewal processes. What is special here is that the probability that an event occurs in an arbitrary interval (t, t+Δt][3] is supposed to be governed by the following three equations (Cox and Miller 1965, page 6, Equations 7-9):

$$p(\text{e occurs once in } (t, t+\Delta t]) = \rho\Delta t + o(\Delta t) \quad \text{(B.16)}$$

$$p(\text{e does not occur in } (t, t+\Delta t]) = 1 - \rho\Delta t + o(\Delta t) \quad \text{(B.17)}$$

$$p(\text{e occurs more than once in } (t, t+\Delta t]) = o(\Delta t) \quad \text{(B.18)}$$

Since $o(\Delta t)$ is by definition a term that is always smaller than Δt, Eq. B.18 means that in an infinitely small interval of time, the event will only occur once. The term ρ in the other two equations, Eq. B.16 and B.17, is called the *intensity* of the process; the higher ρ, the more likely it is that an event occurs. From equations B.16–B.18, it follows that the X_i are exponentially distributed (Cox and Miller 1965, page 147):

Lemma B.1 *Let Z be a variable that corresponds to the time that has elapsed since the last time event e occurred.*
Let the probability that e occurs within an arbitrary interval (t,t+Δ t] conform to equations B.16–B.18.
Finally, let P(x) = p(Z > x) be the probability that the time between the last and the next occurrence of e is greater than x. Then, Z is exponentially distributed:

$$P(x) = e^{-\rho x} \quad \text{(B.19)}$$

Because Z is exponentially distributed, its expectation E(Z) and its variance V(Z) are

$$E(Z) = 1/\rho \quad \text{(B.20)}$$

$$V(Z) = 1/\rho^2 \quad \text{(B.21)}$$

[3]The interval is open at the start time t, and closed at the end time, t+Δt.

So far, we merely have a probabilistic model of the time that elapses between two successive events. Now, we develop this model a little further: we want to predict the number of events that have occurred since the process started. The time S_i that elapses between the first event and the ith event is given by

$$S_i = \sum_{j=1}^{i} X_j \tag{B.22}$$

Because the X_i are exponentially distributed, the mean of S_i is i/ρ, and the variance is i/ρ^2. Now, the number of events $N(t)$ that have taken place from the start time up until a time t is simply the number of events l for which S_l is just below t:

$$N(t) = \max\{l : S_l \leq t\} \tag{B.23}$$

Since the X_i are independent, N(t_1)-N(t_2) and N(t_2) are also independent for arbitrary $t_1, t_2, t_1 > t_2$. N(t) has a Poisson distribution with parameter ρt (for a proof, see Cox and Miller 1965, page 149 f.):

$$P(N(t) = i) = e^{-\rho t}\frac{(\rho t)^i}{i!} \tag{B.24}$$

Therefore the process which describes the counts $N(t)$ is called a *Poisson process:*

Definition B.1 (Stationary Poisson Process) *Let X_i be a series of i.i.d. random variables with $X_i \sim exp(\rho)$ and $S_k := \sum_{i=1}^{k} X_i$. Then,*

$$N(t) = max\{l : S_l \geq t\} \tag{B.25}$$

is a Poisson process with rate ρ.

Poisson processes can be both stationary and non–stationary. The definition I have just given is for stationary processes: For stationary processes, the rate ρ does not change as a function of time. For non–stationary processes, the rate is given by a monotonically increasing, non-negative and continuously differentiable function $\gamma(t)$, which is the same for all X_i. The definition of a non-stationary Poisson process is given below (Leopold 1998, Definition 2.4.2)

Definition B.2 (Non-stationary Poisson process) *A non-stationary Poisson process N(t) is defined by two properties:*

1. *the difference N(t)-N(s), t > s, and the number of events N(s) at time s are independent*
2. $$P(N(t) = i) = e^{-\gamma(t)}\frac{(\gamma(t))^i}{i!} \tag{B.26}$$

   where $\gamma(t)$ is a non–negative, monotonically increasing, and continuously differentiable function with $\gamma'(t)$ bounded.

Definition B.2 may not appear very similar to Definition B.1 for stationary Poisson processes, but in fact, we can adapt it quite easily to the stationary case by replacing $\gamma(t)$ with ρt. A similar definition of stationary Poisson processes can be found in (Resnick 1992, page 303), for more on the relation between the two definitions, see (Leopold 1998, Chapter 2).

B.5.2 Markov Chains

Markov Chains are behind many popular methods in speech and language processing, most notably Hidden Markov Models (Rabiner and Juang 1993). They are also one of the most basic types of stochastic process models.

Markov Chains consist of a sequence of states, which are connected by transitions. The transition probabilities p_{ij} specify the probability of going to state j from state i. For a first-order Markov chain, these probabilities do not depend on the states that have been visited before i; the process is memoryless. Below, I give a more formal definition.

Definition B.3 (Markov Chain) *Let $\{X_i\}$ be a sequence of random variables. X_i specifies the state that a stochastic process is in after the ith transition between states. The initial distribution $\{a_k\}$ specifies for each state k the probability that the process starts in that state:*

$$P[X_0 = k] = a_k, k \geq 0 \tag{B.27}$$

The transition matrix P = $\{p_{ij}\}$ specifies the probability of going to state j from state i in one transition:

$$P[X_{n+1} = j | X_n = i] = p_{ij} \text{ for any } n \geq 0 \tag{B.28}$$

This process is called a **Markov Chain** *iff*

$$P[X_{n+1} = j | X_0 = i_0, X_{n-1} = i_{n-1}, X_n = i] = p_{ij} \tag{B.29}$$

*(***Markov property***)*

A state i is called recurrent iff the process returns to state i in a finite number of steps.

Appendix C The BROWN-COSPEC Corpus

This appendix documents the BROWN-COSPEC corpus that is used extensively in the thesis. Section C.1 describes the annotations and discusses the problem of defining a Genre variable on a representative corpus like Brown, and Section C.2 reproduces the sortal class annotation manual that was used in annotating the texts.

C.1 Annotations of the BROWN-COSPEC Corpus

The Brown corpus of written American English (Francis and Kučera 1979) was designed to be representative of the state of the language at that time. Since then it has become a standard source of data for corpus-based research on American English. These categories have often been taken to correspond to genres. In contrast to text types, which are defined on the basis of text-internal criteria (Biber 1988, Linke, Nussbaumer and Portmann 1994), genres are defined in terms of *non-linguistic* criteria. What these non-linguistic criteria should be depends largely on the researcher. The EAGLES consortium (EAGLES 1996a, EAGLES 1996b) propose as external criteria the participants, the occasion, the social setting, the communicative function of the pieces of language, and so on. While Biber (1988) characterises genres in terms of author/speaker purpose, Swales (1990) defines genres as sets of texts which have a similar communicative purpose in a given discourse community. The norms for genres are set up by the discourse community which produces and reads texts belonging to a genre, and these norms affect both form and content. Genres are inherently fuzzy—some texts are more prototypical for a genre than others. For example, whereas the Gemayel text in Appendix 6.4.3 is a typical newspaper report, the Pratchett text (Appendix A.2) is a parody of the genre it belongs to, fantasy fiction. Recent developments in genre theory combine both aspects of genre: Steen (1999) argues for a prototype theory of genres, where genres that are perceived as prototypical by people are to be listed as bona fide genres and subdivisions of these genres are called sub-genres.

For practical purposes, genre definitions à la Swales are very difficult to transfer to large corpora such as LIMAS or the Brown corpus, because the categories that were used to collect the corpora are often based on content classifications. How can linguists who want to explore language use across genres deal with this problem? Lee (submitted) advocates a pragmatic approach. For him, genres are defined with respect to the *function* of a text, not with respect to a bundle of co-occurring linguistic features. That aspect, the aspect of linguistic *form*, is reserved to register. Lee's function labels are derived from commonly accepted categories and domains, such as social sciences or fiction.

For the LIMAS corpus (Glas 1975), whose categories were taken from the classification

text id	content	source
CF19	argument about American folklore	(Coffin 1961)
CF25	expository, wine-drinking in France	(Churchill 1961)
CF31	argument about the guilt of Lizzie Borden	(Radin 1961)
CG02	expository, history of nationalism	(Miller 1961b)
CG11	argument about the irrationality of fear	(Oppenheimer 1961)
CG35	political speech	(Eisenhower 1961)
CK05	general fiction	(Stacton 1961)
CK25	general fiction	(Miller 1961a)
CK29	general fiction	(Schiller 1961)
CL04	mystery fiction	(Alexander 1961)
CL06	mystery fiction	(Dewey 1961)
CL22	mystery fiction	(Barlow 1961)

Table C.1. Characterisation of texts chosen from the Brown corpus

system of the Deutsche Bibliografie, a library classification scheme which goes mainly by domain, not by genre, Wolters and Kirsten (1999) used classifications which were easy to derive from the existing systematics without having to re-read and re-classify the complete corpus.

In their work on the Brown corpus, Kessler et al. (1997) replaced the existing categories with three *generic facets*, BROW, NARRATIVE, and GENRE. Of the 802 separate texts (from 500 sources) in the corpus, they only chose those 499 which could be classified unequivocally under this scheme. The facets express distinctions that "answer to certain practical interests" (page 33). BROW specifies the "intellectual background" (page 34) that the readers of a text are assumed to have. That this facet also describes to a certain extent the social class of the reader is suggested by the names of the four categories: "popular", "middle", "upper middle", and "high". GENRE corresponds more closely to traditional genres as defined by Swales or Biber. That facet has six values, "reportage", "editorial", "science/technology", "legal", "nonfiction", and "fiction". Finally, NARRATIVE characterises whether a text is a narrative or not. Overall the categories of Kessler et al. (1997) appear to be a mix of external criteria, in particular participant-related, domains (which occur in values of the GENRE facet), and a classical text type, "narrative".

The experiences of Kessler et al. (1997) and Wolters and Kirsten (1999) show that it is difficult to reclassify corpora for genre that have not been balanced properly for that factor in the first place. Lee (submitted) has re-classified all texts from the BNC in great detail. We still need to investigate whether this re-classification scheme can be applied to the Brown corpus and how it would affect our genre categories. But that is definitely future work. For now, we take the pragmatic way out and accepted the genre definitions of the original authors as described in (Francis and Kučera 1979). Table C.1 gives a more detailed overview of the content of the texts we selected for BROWN-COSPEC.

Parts of the Brown corpus have been annotated syntactically in the Penn Treebank project (Marcus, Santorini and Marcinkiewicz 1993). According to the documentation (LDC 1995), the syntactic annotation of the Brown data is the most thoroughly checked of the annotations in the corpus. Therefore it is relatively safe to automatically extract syntactic information from the

Penn Treebank parses of that data. Since corpus annotation can be extremely onerous and time-consuming, existing annotations should be used as far as possible. Therefore we only selected texts for which parses are available in the Penn Treebank version that was used, version 2.0. The twelve texts belong to four groups, CF, CG, CK, and CL. CF and CG are the only non–fiction genres for which parses are available. Only texts were selected which contain very little direct speech.

An extraction program extracted all markables together with information about their form. The texts were annotated further using the annotation tool REFEREE. Both the extraction program and the annotation tool are described by DeCristofaro, Strube and McCoy (1999). The extraction program determined the form of a NP from its Treebank parse: If the POS tag point to a proper name, a personal pronoun, a possessive pronoun, or a demonstrative, the NP is assigned the corresponding value. If the NP contains one of the indefinite determiners "a", "an", or "some", it is labelled as indefinite, if it contains the definite determiner "the", it is labelled as definite. Else, it is assigned the category "none". The implementation we used for pre-tagging our corpora does not distinguish between indefinites and bare NPs; both are lumped together under the category INDEFNP. Subject NPs are labelled in the treebank functional tags. Object NPs are all NP children of a VP. PPs are mostly classified as adjuncts; they are only classified as complements if they have been assigned the functional tag "put" by the Treebank labellers.

After preprocessing, two annotators checked markables and added co-specification sequences, MCUs, agreement labels, and sortal class information. The annotators were first trained on one text. Then each annotator labelled two texts per genre with Sortal Class.

After this brief methodological excursion, let us turn back to the annotation of BROWN-COSPEC. First the annotators checked all NPs that had been marked as referring (in the MUC terminology, *markables*). There were four main types of corrections:

1. unmark reflexives and reciprocals, which are syntactically bound by their antecedents,
2. unmark head NPs of referring expressions whenever the head and the complete referring expression had been analyzed as two separate, nested referring expressions,
3. mark the argument of presentational constructions such as "there was" as referring, and
4. mark constituents of coordinated NPs as referring expressions when they had not been marked separately, and when they were part of a co-specification sequence

In the same pass the annotators labelled co-specification sequences. The only permitted link between a referring expression and its antecedent was identity. The annotation conventions follow those set out in Section 5.2.1. Table 7.5 summarises information about the discourse entities and co-specification sequences in each of the texts and categories, while Table 7.6 presents the frequencies of types of referring expressions.

In a second pass all referring expressions were annotated with agreement and sortal class. Seven agreement classes were defined, combining number and gender features: third person masculine (3M), third person feminine (3F), third person neuter (3N), third person plural (3P), first person singular (1S), first person plural (1P), second person singular (2S), and second person plural (2P). The annotations were based on the surface form. For example, an auctorial pluralis maiestatis was labelled as 1P, although it refers to a single first-person author. Coordinations were labelled as plural, disjunctions as singular. Given that most of these texts were

published in the early Sixties, we chose masculine (M) as the default gender for generic singular references to members of a profession. Incidentally, the third person masculine singular pronoun was also the most frequent choice for pronominal reference to such generic entities. 3F was only used if the persons were referred to were clearly female. The third person impersonal singular pronoun "one" was labelled as 3N.

The sortal class categories are more general than those that were defined for the radio news texts (c.f. Section 6.2.3). As the topics of the BROWN-COSPEC-texts are much more varied than those of the radio news texts, the radio news categories cannot be transferred directly onto the new corpus. The first source of sortal classes that springs to mind are ontologies. Most of the ontologies that have been proposed for natural language understanding systems or lexical knowledge bases have a more or less strict hierarchical structure. If we were to choose all categories from a single level of the ontology, it is likely that this sample would either be too coarse (without a specific category for "person", for example), or too fine-grained, if we pick and choose from several levels of the ontology, we must still make sure that the resulting classes are disjoint. Moreover, there is no such thing as "the" NLP ontology (Guarino 1998, Hovy 1998). While some researchers favour a traditional lexicon-oriented structure (Miller 1995), others decide to develop a full-scale knowledge base (Lenat 1995) which also supports complex inferences from context.

We solved this problem by defining a set of sortal classes on the basis of WordNet (Miller and Fellbaum 1992, Fellbaum 1998). WordNet is a database of lexical semantic relations between English nouns and verbs. Each word meaning is represented by a SynSet, which consists of a short, keyword-like specification of the meaning and links to related SynSets. All WordNet SynSets are linked to one of a number of BaseTypes. BaseTypes are a hierarchically structured set of semantic classes for both noun and verb denotations.

The original class definitions are based on the 71 EuroWordNet BaseTypes (Vossen 1998). These BaseTypes are intended to be language-independent and serve for all language-specific WordNets that were built in that project. The sortal classes are defined extensionally as a set of SynSets. Referring expressions are classified on the basis of their head nouns.

In order to aid annotation, I wrote a small program that determines the potential sortal classes of a noun by looking up hypernymic synsets in WordNet. The initial SynSets that correspond to sortal classes were taken from the definitions of the BaseTypes; some other SynSets had to be added during annotation, when the classification tool failed to find a sortal class that corresponded to the highest hypernymic SynSet of a word. The algorithm is given in Figure C.1. The annotator then has to choose from these sortal classes on the basis of the context in which the word occurs and the descriptions of the sortal classes in the annotation manual, which is reproduced here as Appendix C.2.

In order to establish the reliability of the coding scheme, 8 of the 12 texts were annotated with Sortal Class by two linguists The remaining four texts were annotated by one linguist. The other annotator added MCUs to all texts. [1] The eight texts both annotators had worked on form the basis of the κ-values.

[1] The two annotators were Michael Strube and myself. Michael Strube took over the Sortal Class annotations, while I labelled the units.

Input: head noun of referring expression; pairs of SynSets and sortal classes

1. look up all hypernymic SynSets in WordNet

2. for each meaning m of the word,
   start with synset s corresponding to m
   while ((sortal class not found) or (highest hypernym not reached) do
   if (synset s is associated with a sortal class c) then
   associate meaning s with sortal class c
   flag sortal class found
   associate meaning with sortal class and synset
   elseif (there is a synset s' that is the hypernym of s)
   s := s'
   else
   associate meaning with synset
   flag highest hypernym reached

Output: list of sortal classes and descriptions of SynSets looked up in WordNet.

Figure C.1. Algorithm for computing the potential sortal classes of a word from WordNet hypernymic SynSets

Although there were no labels for uncertain classifications or borderline cases, the annotators could opt for "Sortal Class=none" if none of the classes seemed to fit the referring expression. Since the Sortal Class annotation is quite complex, we measured the reliability of these annotations using Cohen's κ (Cohen 1960, Carletta 1996). A value of κ between 0.68 and 0.80 allows tentative conclusions, while $\kappa > 0.80$ indicates reliable annotations. Overall, the annotations are reliable ($\kappa = 0.8$) Breaking down the results by genre, we find that for CF ($\kappa = 0.83$), CK ($\kappa = 0.84$) and CL ($\kappa = 0.83$), reliability is still fine, but there are problems with genre CG ($\kappa = 0.63$), which contains many abstract discourse entities. Most of the problems are due to the abstract classes Concept, Action, Event, State, and Property. There are two main reasons for this: First some nouns are used metaphorically, and it is up to the annotator to detect the metaphors, second, abstract head nouns sometimes have several senses that fit the context almost equally well, but that lead to different sortal classes.

The division of annotation labour implies that when it came to distilling the final sortal class labels, we could not just take those labels that both of us had agreed on. Instead, I decided to use the judgements of the Sortal Class annotator, except for two texts from genre CL, for which only the second annotator had provided a full annotation. This should not greatly affect our results since the overall agreement was acceptable. In general, for any moderate-scale effort without external funding, such as ours, you need to strike a balance between producing more data or painstakingly checking each annotation where you disagree. If you want to establish gold-standard data sets, the second strategy is clearly right (for recent examples, see e.g. Wiebe, Bruce and O'Hara 1999, Teufel et al. 1999). If you just want enough data for your research, it is sufficient to establish good annotation guidelines and train the annotators well. This appears to be common practice in media studies, where content analysis originated (Früh 1998).

Class	Definition
Person	one or more human beings
Group	institutionalised group of human beings
PhysObj	physical object
Concept	abstract concept
Loc	geographical location
Event	something which takes place in space and time
Action	something which is done
State	state of affairs, feeling, ...
Property	characteristic or attribute of something
Time	date, time span

Table C.2. Overview of Sortal Classes with rough characterisations of relevant synsets

In contrast to the radio news texts the BROWN-COSPEC-texts were not annotated with countability or genericity, because it is very difficult to define a reliable annotation scheme for both of these properties (Poesio, Henschel, Hitzeman and Kibble 1999). For the same reasons, lack of reliable annotation schemes, we neither labeled thematic roles, nor did we classify the predicates of each clause according to time, aspect, and *aktionsart*. This means that the corpus cannot be used to induce algorithms for generating referring expressions or resolving anaphora that rely on such information. For example, in English, genericity substantially influences the choice of determiner (Behrens in preparation, Carlson 1977). If we do not have that information, we cannot predict many of those instances where the definite article may not appear.

C.2 Sortal Class Annotation Manual for the Brown Corpus

C.2.1 Class Definitions

Person: Persons are human beings (Example C.1), metaphorical expressions that refer to human beings such as "life" in Example C.5.
WORDNET BASETYPE: Human
Main synonymous synsets: human 1

Legal/Group: This category includes groups of human beings (Example C.2), institutions (Example C.3), quasi-institutionalized groups of human beings, and companies (Example C.4). We differentiate between Persons and Groups, because this clearly affects the choice of pronouns. It is possible to refer to a group which has been introduced via a NP by a plural pronoun.
WORDNET BASETYPE: Group
Main synonymous synsets: institution, establishment; group

(C.1) **George Bush** is the Republican presidential candidate.

(C.2) **Japanese fishermen** were surprised by the destruction.

(C.3) **The Scripps Institute of Oceanography** is quite famous.

(C.4) **Apple** is a thriving company.

(C.5) **Not a single life** has been lost because of the tsunami.

Time: This category covers temporal expressions. These expressions can refer to days, months, years, days of the week, periods of time, specific points in time, hours and similar subdivisions, as well as expressions such as "the time when the air was still clean".
WORDNET: BaseType Time
Synsets: Time 1

Location: Geographical and astronomical entities such as Japan, Pennsylvania, or Mars are classified as Locations. Locations can be areas, points (as in "the point where . . .") and roads or paths for passage. Buildings and other places such as "my pocket" or "his notebook" were excluded because this would result in a rather fuzzy boundary between Locations and Physical Objects. Some examples are collected in Examples (C.6)–(C.11), locations are printed in boldface, physical objects which are used as locative modifiers in italics.
WORDNET: BaseType Place
Main synonymous synsets: location 1

(C.6) The fishermen in **the Pacific** were surprised by the storm.

(C.7) Earthquakes tell us a lot about **the earth**.

(C.8) **Japan** is an intriguing country.

(C.9) *The hotel room* is clean.

(C.10) *The Empire State Building* is impressive.

(C.11) The student from **Baltimore** had all the formulae in *his notebook*, but not *in his head*.

Physical Object: Physical objects are objects in the real world, such as buildings, heads, computers, trees, waves, and boats. Animals are also categorized as physical objects. We decided against setting up a category "animate" for animals and persons, because personal pronouns referring to persons tend to be marked for male or female gender, those referring to animals not. This is especially pronounced in English.

If the highest hypernymic synset is classed under the WordNet TopConcept "Function", such as possession, product, asset, building, the NP is also assigned the class "Physical Object", because "Function" is only defined for concrete objects.
WORDNET: BaseTypes Inanimate, Animal, and Plant.
Main synonymous synsets: inanimate object 1, material 1, matter 1, animal 1, plant 1, possession 1

(C.12) **The coffee** is still piping hot.

(C.13) Elvis has just left **the building**.

(C.14) **Greebo the cat** is a sly little devil.

(C.15) She always dreamt of **evil mutant killer magnolias** at night.

Event: Events occur in a certain place and at a certain time. An NP should be classified as an event if it can feliticiously replace X in the phrase "X happened". Processes also fall under this heading.
WORDNET: BaseType Event
Main synonymous synsets: event 1, communicate 1, remember 2, consume 2

(C.16) **The conversation** turned unpleasant when the manager mentioned the impending layoffs.

(C.17) The building was completely scorched by **the blazing fire**.

(C.18) After **the ingestion of some fruit**, the animal felt decidedly less hungry.

Action: An action is something that people do or cause to happen.
WORDNET: BaseType Do
Main synonymous synsets: act 1, act 12 (verb), action 1

(C.19) **The hideous murder of Tony Blair** shocked the nation.

(C.20) He has written a Ph.D. about **the ratification of the Yalta pact**.

State: A state is the way something is or a combination of circumstances at a certain time (state of affairs). Feelings and general psychological features are also states.
WORDNET: BaseType State
Main synonymous synsets: situation 1, be 4 (verb), state 1

(C.21) **The love affair between the professor and her secretary** shocked the department.

(C.22) Doctors deal with **illnesses** every day.

(C.23) Words cannot express **the hate** he felt for him.

Property: This category covers expressions which refer to a property of a human being, a concept, or a physical object, such as length, understandability, velocity, or curiosity. Properties are attributes and especially characteristic attributes. NPs whose head is a unit of measure count as **properties**, as well, because they express values of quantifiable properties.
WORDNET: BaseType Property
Main synonymous synsets: property 2, attribute 1

(C.24) **The speed of the boat** was what made them suspicious at first.

(C.25) Many people don't like **the colour red**.

Concept: Concepts are abstract or general ideas inferred or derived from specific instances. For example, concepts can be formed by extracting common features from examples. This is the WordNet definition of "abstraction". The category "concept" is also used for labelling NPs which refer to the contents of cognition or the focus of somebody's thoughts. The category also covers concepts such as information and communication. Finally, feelings and mental attitudes are categorized as concepts.
WORDNET: BaseType Concept
Main synonymous synsets: concept 1, abstraction 1, cognitive content 1, idea 2, communication 1, information 1, cognition 1, feeling

(C.26) **The information** was extremely valuable to the spies.

(C.27) **The sum** of everything that has been said is that money stinks.

(C.28) **Words** cannot express the hate he felt for him.

Not Classifiable: This category is a repository for all anaphoric expressions that do not fall in one of the categories sketched above.

C.2.2 Annotation Strategy

Ideally, we would be able to feed WordNet the NP to be classified, and out comes the correct ontological class. For several reasons, this is not practical:

- WordNet does not know all possible names of persons. Neither can it detect names of institutions which do not contain a cue word such as "university", "institution" or "group". Thus, the concepts "person" and "group" have to be labelled by hand.
- Locations which are referred to by their proper names and dates which are referred to numerically are also difficult to look up. However, given the definition, locations and times can be determined manually.
- A word usually has multiple senses, and the adequate sense has to be determined manually.

To overcome these problems, we propose the following strategy:

1. Before proceeding to sortal class markup, the referring expressions should have been labelled for coreference, agreement, and syntactic function.
2. The markup then proceeds one coreference chain at a time. First, it is checked whether the referent of this chain is of a type that would present difficulties to WordNet lookup - Person, Group, Location, or Time.

   If yes, it is classified according to the decision tree in Figure C.2 and the definitions in Section C.2.1.

   If not, the hyperonyms of the head noun of one of the referring expressions are looked up in WordNet. If there are multiple senses, the adequate sense is selected. For this sense, the

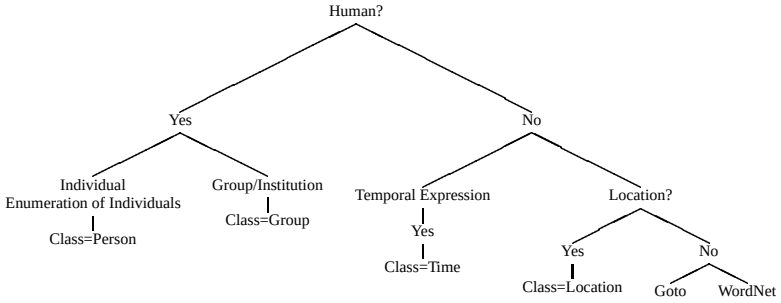


Figure C.2. Decision Tree for labelling categories where WordNet is inherently unreliable.

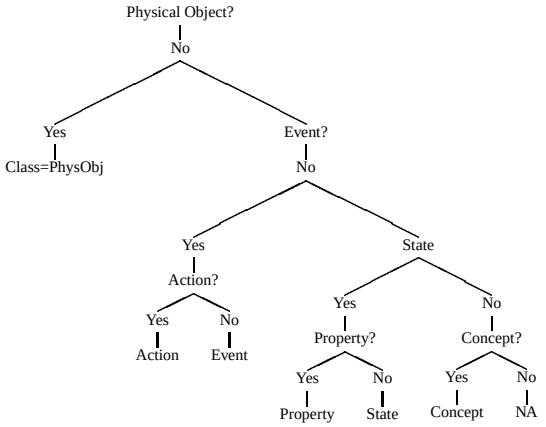


Figure C.3. Decision Tree for WordNet-Classification. It presents the order in which categories should be checked if WordNet does not contain the word queried.

list of hyperonyms is searched for a synset that was mentioned in Section C.2.1 as one of the main synonymous synsets of a category. If there are several potential candidates, the most specific one is selected. The NP is then assigned the sortal class which is associated with that synset.

If the category of the head noun does not appear to match the sortal class of the complete NP, the NP has to be classified manually according to the decision tree in Figure C.3.

Appendix D Ungeheuer's Approach to Communication

In this appendix, I discuss Gerold Ungeheuer's approach to communication in some more detail, because his work is relatively little-known, even in Germany. In fact, Ungeheuer has never published a full-blown theory of communication, and his writings are fragmentarian. But the assumptions, hypotheses and concepts that Ungeheuer develops in these fragments are powerful tools for thinking about communication — both on the micro- and on the macrostructural level. The aim of this outline is to present some of these conceptual tools which I will use for developing a more general concept of entity status in communication. For this reason, I will not to discuss the philosophical and semiotic background of Ungeheuer's approach in detail, or to dwell on the work of those that his work is in turn based on, such as Karl Bühler, Philipp Wegener, Alfred Schütz, Max Scheler, or Stephen Toulmin, to name but a few.

In a nutshell, Ungeheuer conceives of communication as a process in which both parties, speaker and hearer, take an active part. A speaker intentionally performs actions in order to communicate something to a hearer, and the hearer interprets these actions so that he can reconstruct what the speaker wanted to communicate to him. Both parties act on the basis of their individual theories of the world. Thus, Ungeheuer sees communication as a process of social inter*action*.

This approach differs markedly from two other approaches that have been far more popular in the literature: (Watzlawick et al. 1967) and (Shannon and Weaver 1949).

In the classic Shannon-Weaver model (Shannon and Weaver 1949), a sender encodes a message, the message is transmitted via a channel to the receiver, and the receiver decodes that message. Communication fundamentally means that information has been transmitted. In Ungeheuer's approach, on the contrary, communication is fundamentally asymmetric. The speaker intends to influence the hearer so that the hearer performs certain internal actions. The speaker does not seek to get some message transported to the hearer, she seeks to exert control over him, to get him to do something.[1] The internal actions which are performed when producing and interpreting a sequence of signs are highly structured and complex. Signs are not fixed codes with specific information values. Rather, the meaning of a sign is always constructed anew, depending on who interprets it in which situation.

Watzlawick et al. (1967) claim that it is impossible not to communicate, because every aspect of your behaviour can be interpreted as a sign for something. Ungeheuer, on the contrary, stresses that people can choose to communicate, and that communication is intentional. If

[1] Ungeheuer elaborates this view on the basis of a detailed analysis of (Bühler 1927/1965) in (Ungeheuer 1967/1972a).

your addressee does what you intended, you are successful — although it is very difficult to determine whether that success has indeed taken place. If, on the other hand, the addressee mistakenly assumes some part of your behaviour, verbal or non-verbal, as a sign, this does not mean that you intended this behaviour to be a sign. You did not communicate, because you did not intend to perform an action that gets the other to interpret that action as a sign. For example, I might speak to you in a monotonous voice. You may interpret this as a sign that I am bored stiff by you, because somebody else has already used this trick on you before. In fact, it may be that I am just tired, and this is reflected in my voice.

Let us now explore the premises behind Ungeheuer's approach to communication in more detail. For Ungeheuer, communication is an action. Actions are intentional behaviour. More precisely, it is an indirect or mediated interaction between humans, which implies that the medium of communication is not necessarily language:[2]

> Indirekte oder vermittelte Interaktion zwischen menschlichen Individuen sei *Kommunikation* genannt.[3]
>
> (emphasis in the original Ungeheuer 1974/1987b, page 83)

Interactions, in turn, are social actions. They occur when somebody aims to make another person perform a specific action, and performs an action himself in order to reach that goal.

> [. . .] "Interaktion": [. . .] man führt eine Handlung aus mit dem Ziel, das andere Individuum zu bestimmter Handlung zu bringen.[4] (Ungeheuer 1974/1987b, page 82)

Communication can be embedded in other social actions as a means of reaching the goal of those actions.

But what exactly are those actions that Ungeheuer refers to, and why does he insist on using this concept? First of all, actions are *intentional*. Actions are performed with a goal in mind. In communication processes, that goal is to make oneself understood, in Ungeheuer's terms *Verständigung*. Ungeheuer's concept of social action owes much to Weber and Schütz (key publications: (Schütz 1960, Weber 1940)). For Weber, social actions are those actions whose goal is defined in terms of the behaviour of others. Schütz emphasises that it is only possible for actions to make sense if one thinks about them "modo futuri exacti" (Schütz 1960, page 60), as if they had already happened. In order to plan an action, the actor has to imagine the result she would like to achieve with it.

But what does a speaker need to achieve if she wants to make herself understood? Ungeheuer answers this question in a paraphrase: When I speak, I want the person I speak with to understand me (Ungeheuer 1974/1987e, page 34). Understanding is again an action, but this time an internal action. The dichotomy between internal and external aspects of actions is a long-standing problem in action theory (Connolly 1989, Luckmann 1988), and Ungeheuer's

[2] All English translations of the original quotes are intended to help the reader understand what Ungeheuer intended to say; they are not authoritative.

[3] Let us call indirect or mediated interaction between human individuals *communication*.

[4] "interaction": an action is performed with the aim of getting the other individual to perform a certain action.

approach to it owes much to scholars such as Schütz (1960). Actions are intentional behaviour. Behaviour can be observed, it is external, but intentions cannot. An external observer can speculate about the intentions of the person who performs an action on the basis of social conventions, on the basis of what that person told her, on the basis of her own experience, and on the basis of her experiences with that person – but not more. Ungeheuer distinguishes between two types of actions, *external actions* which can be perceived by the sensory organs, and *internal actions* which can only be perceived by the person who acts.

> a) äußere Handlungen von Menschen sind solche, die durch Sinnesorgane wahrnehmbar sind;
> b) innere Handlungen von Menschen sind solche, die nicht durch Sinnesorgane wahrnehmbar sind, - sie sind direkt erfahrbar nur dem so handelnden Individuum selbst.[5]
>
> (Ungeheuer 1974/1987e, page 41)

Luckmann (1988, Kurseinheit 2,page 8) distinguishes between an external and an internal perspective on action (*"Außen"- und "Innen"-Perspektive des Handelns*). The external perspective is that of the observer, who can only experience actions as mediated by behaviour, and the internal perspective is that of the actor, who experiences his act immediately. Only external actions can be seen from an external perspective. All internal actions are only accessible to the actor herself. In communication, the external action is the sequence of symbols that the speaker produces. The actions which take place when that sequence is planned by the speaker, and interpreted by the hearer, are internal. The following quote summarises the role of internal actions in communication:

> In [der Kommunikation] wird die Handlung, die Ziel der Interaktion (als Sozialhandlung) ist, durch eine Zwischenhandlung erreicht, in der die Verwirklichung des Handlungsziels nicht nur für den Initianten, sondern auch für den Akzeptanten der Interaktion durch *kognitiv–hypothetische Vorwegnahme* vermittelt ist.[6]
>
> (Ungeheuer 1974/1987b, page 83)

Let me illustrate this with an example: Suppose that I want you to know that Germany lost 3:0 to Portugal. My goal is that you add this to what you know about the world. I now plan how to perform that action, and decide that the best option is to utter the following: "Oh by the way, the German soccer team lost 0:3 to Portugal." You hear my utterance and try to interpret what I intended with this action. If you think that I just want to claim your attention by uttering some nonsense, if you think something like "How on earth could the Portugese beat the European Champion Germany 3:0? She must be crazy! No, it must be the other way around:

[5] a) external actions of humans are those that can be perceived via the sensory organs;
b) internal actions of humans are those that cannot be perceived via the sensory organs, - they can only be perceived directly by the very individual who performs them.

[6] In [communication], the action which is the goal of the interaction (as social action), is reached by an intermediary action, in which the realisation of the aim of the action is mediated by *cognitive-hypothetical anticipation* not only for the initiator, but also fro the acceptant of the interaction.

Portugal:Germany 0:3.", you will draw conclusions about me, but you will not add the fact that Portugal beat Germany to your knowledge, and I was unsuccessful. If you take me seriously, however, you believe me, and my action had the intended consequences: you added the fact that Germany lost 3:0 to Portugal.

In Ungeheuer's view, all linguistic signs (and hence also all discourses that consist of these signs) are instructions to perform actions. This view is explained in more detail in his last paper, (Ungeheuer 1982/1987d), where he explicitly states the assumptions behind his approach to communication. The actions that the hearer is instructed to perform are internal. This means that the speaker cannot observe directly whether the hearer has indeed performed the action she wanted him to perform. I cannot open your mind and search for the place where you have stored Germany:Portugal 0:3, I can only hypothesise that you have done so. In this action-theoretic perspective, the hearer has three tasks:

- decide which parts of the speaker's behaviour is to be interpreted as an action $\rightarrow$ as a sign
- determine which internal actions the speaker intended him to perform
- perform these internal actions on the basis of his personal experience theory (PET), if he decides to. More often than not, he will do that without even reflecting on what he is doing, but in states of heightened consciousness, for example when doing Critical Discourse Analysis, he can reflect on these actions

A possible objection against this action–theoretic approach to language might be that since actions are intentional, they should always be conscious. The sequence of events that I have outlined above appears to be too laborious to be behind everyday speech. But actions can be more or less routine, more or less deliberately planned and executed. Luckmann (1988) cites the example of walking. Usually a highly automatic action, it has to be planned and executed very slowly and deliberately when the person who wants to walk has temporarily lost the use of her legs, until it becomes routine again. Speech is also highly routine. The actions that are associated with producing and interpreting verbal signs are largely conventionalised. This explains why we are often not conscious of the complex actions that lie behind our everyday talking. Only when we fail or when we very carefully want to avoid failure do we become aware of this.

The aim of the addressee's internal actions is to experience that he understands the sequence of signs that the communicator has produced. When he has understood that sequence, he has gained new knowledge, or he has linked something which he already knows. Ungeheuer summarizes this train of thoughts as follows:

> H3: In kommunikativer Sozialhandlung sind Formulierungen und Teilformulierungen bis zu jedem Sprachzeichen Anweisungen und Pläne für den Hörer zum Vollzug von inneren Erfahrungsakten, von denen der Sprecher annimmt, ihnen würden Inhalte korrelieren, die er meint.[7] (Ungeheuer 1982/1987d, page 316)

[7] In communicative social actions, the formulations and partial formulations up until every linguistic sign are instructions and plans for the hearer to perform internal acts of experiencing. The speaker assumes of these acts that those contents would be correlated with them which he means.

But these gains and links are not necessarily those that the communicator wanted the hearer to make. In fact, as a communicator, I have to assume that I can make you, the addressee, understand the sequence of signs I produce in the way I intend you to understand them. Else, I cannot communicate with you. While I communicate with you, you must allow me to instruct you how to understand me, and you must try to follow these instructions. Both partners need to cooperate if communication is to be successful, but the communicator leads, and the addressee has to follow. This is what Ungeheuer calls *kommunikative Subjektion*, communicative subjection.

To cooperate in communication is difficult. This is a common experience, and will not surprise anybody. But why these problems? Because the goal of the communicator is to make the addressee perform internal actions, and since she cannot manipulate his mind directly, she needs to resort to external actions, actions that can be observed. External actions are also the only way to check whether a desired action has been performed by the addressee. If communication is only part of a larger social action, then the success of that communication can be judged by the success or failure of (parts of) that action. This criterion is close in spirit to the criterion that (Brown 1995) uses in experimental studies of communication. Alternatively, the communicator can try to infer from external actions of the addressee that he has understood her; if necessary, both can then negotiate problematic points.

The communicator can only plan her actions based on what she knows. Her knowledge can include experiences with and hypotheses about her addressee, but that will never be fully adequate, since she cannot read his mind. The experiences he makes when he interprets her signs are internal. Her decisions are based on her *individuelle Welttheorie*, her individual theory of the world, or personal experience theory (PET).[8]

> Das [. . .] vielgliedrige und in der ständigen Bewegung des Auf- und Abbaus sich befindliche, manchmal in mir als strömend erlebte Erfahrungssystem, das ich bin, nenne ich in begrifflicher Repräsentation meine *individuelle Welttheorie*.[9]
>
> (Ungeheuer 1982/1987d, page 312)

My individual theory of the world determines how I experience something. What controls my experiences is a collection of emotions, intuitions, systematic and unsystematic thoughts, assumptions. Complex rules govern how that collection grows and changes. This collection, which is completely internal to me, constitutes a dynamic system of experiences. In the quote, Ungeheuer calls this system "the system of experiences that I am (*Erfahrungssystem, das ich bin*" (1982/1987d, page 312)). My individual theory of the world also determines how I communicate: what effect I would expect the signs I produce to have on my addressee, and how I would interpret signs that come from another person.

So far, I have sketched the basic tenets of Ungeheuer's approach to communication. The picture that emerges is dynamic, not static. This is the result when we describe communication in terms of actions. Both speaker and hearer change in the process of communication. Their system of experiences, their individual theory of the world, is changed by their exchange. Signs

[8] Ungeheuer chooses the term "theory", because the individual theory of the world determines how experiences are explained.

[9] The [. . .] multifaceted, perpetually being built and torn apart, sometimes experienced as flowing, system of experiences that I am, is what I call in conceptual representation my *personal experience theory*.

are no static entities, either. Every time a sign is used in communication, the hearer constructs an interpretation on the basis of his current system of experiences, which changes continuously in time, and every use of a sign in communication potentially influences the way in which it will be used the next time by the participants in the communication process.

So far, we have defined communication as a mediated interaction between communicator and addressee. The communicator wants the addressee to perform certain internal actions. Both have different theories of the world, which influence how they use signs. On the basis of this definition, let us now turn to that which is being communicated, the content. We will concentrate on linguistic communication here. Ungeheuer distinguishes between the primary, secondary, and tertiary content (Ungeheuer 1967/1972a) of a discourse.[10] The primary content is what the discourse is about, the central train of thoughts. It has two main components, material and modal. The material component is very roughly the semantic content of an utterance. This content has not yet been organized in terms of psychological subjects and psychological predicates.[11] Such partitions come into play when the material component is adapted to a communication situation, with specific addressees, attitudes, and functions that the utterance has to fulfil in discourse. Such functions can be specified by argumentation schemes, dialogue acts, etc. The list of such situation-dependent modifications is potentially infinite. Taken together, they constitute the modal component. Ideally, the addressee should follow the instructions in both components.

The secondary and tertiary content are both constructed actively by the addressee, and are never mentioned explicitly in the discourse. The secondary content consists of that which the addressee infers about the communicator on the basis of the discourse, while the tertiary content is derived from the discourse when the addressee associates what he knows from the discourse with what he already knew before, but what has not been mentioned explicitly in the text.

Ungeheuer did not develop the categories of primary, secondary, and tertiary content much further in his research. In his published writings, he focussed instead on philosophical, semiotic, and epistemological problems. Whenever he demonstrates how the approach he developed can be used for analysing of conversations, he tends to discuss the primary component in great detail, describing it in terms of categories which he drew from the theory of argumentation (Ungeheuer 1972c, Ungeheuer 1974/1987b, Ungeheuer 1980/1987a). To facilitate his analysis, he paraphrases the excerpt to be analysed so that primary content emerges clearly. This is also the strategy that Hanke (1984) uses for analyzing the structure of university lessons.

To readers with a computational or computational linguistics background, Ungeheuer's terminology, especially in this brief summary, may appear to fit nicely into AI frameworks where cognitive models are constructed and manipulated, where all plans are rational, and where agents try to maximize their expected utility. This is a misunderstanding. Ungeheuer's communicators and addressees are not robots. They use heuristics, they make mistaken assumptions, they decide emotionally, they think of weird "plans" for actions which would be rejected by any sane computer program. Juchem (1998) has shown that emotions are indeed a possible, plausible, and even necessary part of a theory that is based on Ungeheuer's approach.

[10] His original term is "text", because he developed these concepts in a discussion of content analysis. We substitute discourse here, because that term is less laden with connotations of written language.

[11] Other common terms in the literature for that dichotomy are theme/rheme, topic/focus, and focus/background. The subject/predicate terminology can be traced back to (Paul 1920, von der Gabelentz 1891).

Anhang E Zusammenfassung

Dieses Kapitel enthält eine informelle Zusammenfassung der Arbeit in deutscher Sprache. Jedem Abschnitt entspricht ein Kapitel im englischsprachigen Hauptteil. Die Einleitung (Abschnitt E.1) wurde für diese Zusammenfassung neu geschrieben; die anderen Abschnitte sind eng an die Zusammenfassungen aus dem englischsprachigen Teil der Arbeit angelehnt.

E.1 Einleitung

Ausgangspunkt der Arbeit war die Beschäftigung mit dem linguistischen Begriff der “Givenness”, also der Gegebenheit von Informationen im Diskurs, mit dem in der Literatur sehr gerne argumentiert wird. Ziel dieser Arbeit war es, diesen Begriff so einzuschränken und zu beschreiben, dass man mit ihm vernünftig arbeiten kann. Das ist ein nicht ganz neues Anliegen in der Linguistik, wie die ironischen Kommentare von Prince (1981) zeigen. Und wie der Prolog zu dieser Arbeit bereits andeutet, glaube ich nicht, den Stein der Weisen gefunden zu haben. Was ich im sogenannten Theorieteil dieser Arbeit beschreibe, ist eine Systematisierung, die mir bei meiner eigenen Arbeit geholfen hat, die andere Forscherinnen und Forscher vielleicht auch hilfreich finden könnten, und von

der ich annehme, dass sie mehr empirischen Gehalt besitzt als die Theorie über Brontosaurier, die im Epilog so eloquent vorgetragen wird—ganz abgesehen davon, dass ich meinen Beitrag nicht als Theorie bezeichnen würde.

In dieser Arbeit habe ich einen ähnlichen Weg beschritten wie schon viele andere vor mir, die oben erwähnte Ellen Prince eingeschlossen: Ich habe mich darauf beschränkt, Diskursentitäten zu betrachten. Diskursentitäten sind Hilfskonstrukte zur Interpretation von Diskursen; sie stellen Platzhalter dar, auf die anaphorisch zurückverwiesen und an denen Information aufgehängt werden kann.

Im Gegensatz zu meinen Vorgängern auf diesem ausgetretenen Pfad entschied ich mich, meine Erkenntnisse nicht an eine bestimmte linguistische Richtung zu binden, wie etwa die Richtung der Optimalitätstheorie (Prince and Smolensky 1993) oder der systemisch-funktionalen Grammatik (Halliday 1994). Ich bin überzeugt, dass in der Linguistik nur der Methodenpluralismus weiterführt. Da es den Rahmen einer Doktorarbeit, einer Textsorte, an die gewisse minimale Kohärenzanforderungen gestellt werden, sprengen würde, mehrere Theorien zu entwickeln und zu vergleichen, habe ich mich für den kleinsten gemeinsamen Nenner entschieden: die Exploration des Gegenstandes, der untersucht werden soll. Um terminologische Verwirrungen zu vermeiden, habe ich diesen Gegenstand “Entitätenstatus” genannt. Der Status einer Diskursentität ist ein Hilfskonstrukt. Er beschreibt, welche Rolle diese Entität im Diskurs spielt (*Strukturdimension*), und stellt Informationen für diejenigen Prozeduren bereit,

die diese Entität verwalten (*Verwaltungsdimension*).

Die Arbeit ist in acht Kapitel und fünf Anhänge (einschliesslich dieser Zusammenfassung) gegliedert. Die Kapitel entsprechen den Abschnitten dieser Zusammenfassung; die Anhänge werden, soweit sie wichtige Ergebnisse enthalten, im Rahmen der betreffenden Kapitel diskutiert.

Im zweiten Kapitel (Zusammenfassung in Abschnitt E.2) erforsche ich den Begriff der Diskursentität, stelle ihn in seinen historischen Kontext, und schlage eine semiotische Deutung vor. Darauf aufbauend definiere ich, was ich unter dem Status dieser Diskursentitäten verstehen werde. Schliesslich diskutiere ich sowohl Diskursentitäten als auch ihren Status unter kommunikationstheoretischer Perspektive. Diesen für eine computerlinguistische Arbeit ungewöhnlichen Schritt habe ich gewählt, da Computerlinguisten bei der Modellierung immer Abstriche machen müssen. Mich interessierte, was dabei herauskommt, wenn man sich der Komplexität des Gegenstandsbereiches einmal furchtlos stellt, und wie man am geschicktesten seine Rückzugsmanöver vorbereitet, wenn man sieht, dass dieser Komplexität mit computerlinguistischen Mitteln nicht Herr zu werden ist.

Im Kapitel 3, zusammengefasst in Abschnitt E.3, erforsche ich, wie man beschreiben kann, welche Rolle eine Diskursentität in einem Diskurs spielt. Dabei konzentriere ich mich auf Theorien, die in der Computerlinguistik sowie in der angloamerikanischen Linguistik eine grosse Rolle spielen. Drei grosse Themenbereiche werden diskutiert: Kohärenz und Kohäsion, Diskursstruktur,

und Thematizität.

Kapitel 4 (zusammengefasst in Abschnitt E.4) ist der Verwaltung von Diskursentitäten gewidmet. In diesem Kapitel beschreibe ich zunächst ein kleines Bestiarium anaphorischer Ausdrücke, um daraus abzuleiten, was eine Theorie abdecken muss, die erklären will, wie Diskursentitäten verwaltet werden. Dann gehe ich näher auf eine Forschungsrichtung ein, die sich ebenfalls damit beschäftigt hat, wie Menschen auf Diskursentitäten verweisen können, und auf die in der linguistischen Literatur dementsprechend oft verwiesen wird: auf die Psycholinguistik. Zum Schluss diskutiere ich, wie Linguisten mit dem Problem des Zugriffs auf Diskursentitäten umgegangen sind. Ich konzentriere mich dabei auf die vielzitierten Arbeiten von Prince (1981), Lambrecht (1994), Grosz et al. (1995), Ariel (1990), Gundel et al. (1993), Givón (1995a), and Chafe (1994).

Damit ist der Theorie-Teil, in dem ich wichtige Arbeiten und Forschungsrichtungen zum Thema Entitätenstatus aufarbeitete, abgeschlossen. Das nun folgende Kapitel 5 (zusammengefasst in Abschnitt E.5) legt die methodologische Grundlagen für den folgenden empirischen Teil. Zunächst diskutiere ich kritisch einige Annotationsschemata für Koreferenz, oder, wie ich hier in Anlehnung an Webber (1983) und Sidner (1983) sagen werde, Kospezifikation. Auf der Grundlage dieser Schemata entwickele ich ein eigenes Annotationsschema, das zur Bearbeitung der Radionachrichtentexte eingesetzt wurde, die in

Kapitel 6 ausführlich analysiert werden. Dieses Schema klassifiziert referierende Ausdrücke danach, woher die Diskursentität stammt, auf die sie verweisen bzw. (um in der Terminologie zu bleiben) die sie spezifizieren. Zum Schluss dieses Kapitels denke ich darüber nach, was man tun kann, wenn es sich als unmöglich erweisen sollte, Texte zuverlässig mit solch einem quellenbasierten Schema zu annotieren. Dann kann man im Grunde genommen nur noch auf Maße zurückgreifen, die sich anhand vorannotierter Kospezifikationsfolgen berechnen lassen. Nach einer ausführlichen Diskussion der verschiedenen Alternativen stelle ich ein solches Abstandsmaß vor. Auf der Grundlage dieses Abstandsmaßes skizziere ich ein statistisches Modell, das die Abfolge von Erwähnungen einer Diskursentität in einem Text beschreibt. Da die mir vorliegenden Korpora nicht gross genug sind, war es nicht möglich, dieses Modell empirisch zu verfeinern und zu testen.

Im folgenden empirischen Teil prüfe ich, welche Erkenntnisse sich über den Status von Diskursentitäten aus annotierten Korpora gewinnen lassen, wenn man die in Kapitel 5 vorgestellten Annotationstechniken verwendet.

Kapitel 6, zusammengefasst in Abschnitt E.6, ist einer ganz besonderen Textsorte gewidmet: den Radionachrichten. Im ersten Abschnitt dieses Kapitels ergründe ich, was Radionachrichten als Textsorte so besonders macht, und worauf man bei der Analyse dieser Textsorte achten sollte. Danach stelle ich die Daten vor, auf denen ich gearbeitet habe, und beschreibe die Annotationen, mit denen ich sie angereichert habe. Diese Daten bilden die Grundlage für die

folgende, eingehende quantitative Analyse. Als Gegengewicht zur quantitativen Analyse folgt eine qualitative Analyse von fünf Texten, einem Zeitungstext und vier Radionachrichtentexten aus dem DLF-Korpus. Der Schwerpunkt dieser Analyse liegt auf Einflüssen auf den Entitätenstatus, die sich nur schwer annotieren und quantifizieren lassen, wie z.B. der Kontext, in dem eine Nachricht geschrieben wurde, sowie ihr Nachrichtenwert.

Für Kapitel 7 bestand das Datenmaterial aus einem mit Kospezifikationsfolgen annotierten Subkorpus des Brown-Korpus (Francis und Kučera 1967). Dieses Subkorpus, das ich im Folgenden als BROWN-COSPEC bezeichnen werde, besteht aus Texten verschiedener Textsorten. Es ist in Anhang C dokumentiert. Dieser Anhang beschreibt Struktur und Inhalt des Korpus, diskutiert in diesem Zusammenhang kurz die Textsortenproblematik bei so genannten repräsentativen Korpora und gibt das Annotationshandbuch für die Sorten-Ontologie wieder.

Zum Schluss untersuchen wir in Kapitel 7, zusammengefasst in Abschnitt E.7, Pronominalisierungsmuster im BROWN-COSPEC-Korpus. Meine Leitfrage dabei ist—und diese Leitfrage differiert von der in (Strube und Wolters 2000)—wie wichtig Entitätenstatus ist (hier: strukturelle Dimension, gemessen in Abstand zur letzten Erwähnung), um zu erklären, wann ein referierender Ausdruck als Pronomen, und wann er als volle NP realisiert wird. Dazu führten wir detaillierte statistische Analysen des Korpus mit Hilfe der logistischen Regression

durch, einer Analysetechnik, die es erlaubt, genaue “Anfragen” in Hypothesenform an ein Korpus zu stellen. Die gewonnenen Ergebnisse haben wir anhand von Experimenten überprüft, bei denen ein Regelinduktionsalgorithmus (RIPPER, Cohen 1995) und ein exemplarbasierter Lerner (IBL-IG, Daelemans et al. 1997) anhand der Korpusdaten lernen sollten, wann zu pronominalisieren ist, und wann nicht. Bei unseren Ergebnissen richteten wir besondere Aufmerksamkeit auf Genreunterschiede: Wir waren an Faktoren interessiert, die Pronominalisierung in allen Genres etwa gleich gut vorhersagen.

Kapitel 8 greift die wichtigsten Erkenntnisse der Arbeit noch einmal auf und diskutiert die Frage, mit der alles begann: Was bringt eine Analysekategorie “Gegebenheit” für die praktische linguistische Arbeit?

Anhang A gibt zwei Texte wieder, auf die im Verlauf der Arbeit des öfteren verwiesen wird: einen kurzen Zeitungstext über die Ermordung des libanesischen Präsidenten Bashir Gemayel Anfang der achtziger Jahre des letzten Jahrhunderts, und den Beginn des Romans “Guards, Guards” von Terry Pratchett. Anhang B führt kurz in einige Methoden ein, die in Kapitel 6 und Kapitel 7 verwendet werden, aber in einigen Zweigen der Linguistik nicht allgemein bekannt sind: allgemeine lineare Modelle und statistische Prozesse. Anhang C beschreibt das BROWN-COSPEC-Korpus, und Anhang D führt in Gerold Ungeheuers Kommunikationstheorie ein.

E.2 Was ist Entitätenstatus?

Viele Geisteswissenschaftlerinnen und Geisteswissenschaftler, und auch einige Naturwissenschaftlerinnen und Naturwissenschaftler, nehmen an, dass es nicht möglich ist, ein umfassendes, mathematisch sauberes Modell der menschlichen Kommunikation zu entwickeln. Diesem Standpunkt schliesse ich mich an. Ich gehe jedoch ebenfalls davon aus, dass man formale Methoden braucht, um Muster und Strukturen menschlicher Kommunikation präzise und knapp zu beschreiben—das heisst, diejenigen, die sich halbwegs objektiv beobachten lassen. Da die meisten Ergebnisse, die ich in den späteren Kapiteln dieser Arbeit berichten werde, quantitativer Art sind, also statistische Modelle und statistische Analysen beobachteter Daten, brauche ich zum Ausgleich eine Perspektive, die sehr genau die *Grenzen* eines solchen Ansatzes aufzeigt, die unterstreicht, was wir verlieren, wenn wir anfangen zu zählen und aufhören zu interpretieren.

Deshalb habe ich in Kapitel 2 eine etwas ungewöhnliche Perspektive eingenommen: die Perspektive der Semiotik und der Kommunikationstheorie nach Gerold Ungeheuer. Ungeheuer betrachtet Kommunikation als Prozess. An diesem Prozess sind beide, Sprecher wie Hörer, aktiv beteiligt. Ein Sprecher führt beabsichtigte Handlungen aus, um etwas einem Hörer mitzuteilen, und der Hörer interpretiert diese Handlungen, damit er rekonstruieren kann, was der Sprecher ihm mitteilen wollte. Beide Kommunikationsteilnehmer können nur auf Grundlage ihrer individuellen Welttheorie, im folgenden “persönliche

Erfahrungstheorie" oder kurz PET genannt, handeln; sie können die Erfahrungstheorie des anderen nie erkennen, sondern nur ihre Schlüsse aus dessen beobachtbarem Verhalten ziehen. Anhang D führt kurz in Gerold Ungeheuers Theorien ein. Weitere Einführungen finden sich in (Juchem 1989) oder (Juchem 1998) passim, der Ungeheuers Gedankengut in Richtung des Radikalen Konstruktivismus weiterentwickelt hat, einer Richtung, die ebenfalls in der Medienwissenschaft stark vertreten ist.

Aus dem Blickwinkel einer dem Werk von Charles Sanders Peirce verpflichteten Semiotik ist es sehr sinnvoll, bei der Interpretation von referierenden Ausdrücken zwischen einem Referenten und einer Diskursentität zu unterscheiden. Denn jedes Zeichen ist verbunden mit einem Objekt in der Welt (hier: dem Referenten) und einem Interpretanten (hier: der Diskursentität), der dieses Zeichen interpretierbar macht, mit anderen Zeichen verbindet, und damit in den unendlichen Prozess der Semiose einführt. Diese semiotische Interpretation ist eine etwas ungewöhnliche Umdeutung der normalen Definition von Diskursentität als mentaler Repräsentation dessen, worauf sich anaphorische Ausdrücke zurückbeziehen können (Sidner), oder einer Beschreibung als "konzeptueller Kleiderhaken" (Woods, zitiert in einigen Arbeiten von Webber).

Aus einem kommunikativen Blickwinkel, also aus einer Perspektive, die sowenig wie möglich betrachtend und abstrahierend ausserhalb des Kommunikationsprozesses steht (Ungeheuer 1967/1972a), sind Diskursentitäten die "extrakommunikativen" Formalisierungen von Einheiten im Fluss der Erfahrungen,

die Sprecher und Hörer im Verlauf eines Diskurses machen. Die Grenzen dieser Einheiten sind nicht scharf, und sie sind eingebettet in das Netz von Erfahrungen, das die PET jedes Menschen darstellt. Die Einheiten sind dynamisch: jedesmal, wenn sie zum Verständnis eines Diskurses benötigt werden, ändert sich das Erfahrungsnetz, an dem sie teilhaben, leicht.

Der Status einer Diskursentität kann entlang zweier Dimensionen definiert werden:

die Struktur-Dimension: diese Dimension beschreibt die Rolle, die eine Entität in dem Teil des Diskurses spielt, in dem sie erwähnt wird und

die Verwaltungsdimension: diese Dimension beschreibt die Informationen, die bereitgestellt sein müssen, damit eine Diskursentität vernünftig initialisiert werden kann, damit auf diese Entität zugegriffen werden kann, und damit die Beschreibung dieser Entität vernünftig aktualisiert werden kann.

Die beiden Dimensionen überlappen einander. Je zentraler die Rolle ist, die eine Diskursentität in einem Diskurs spielt, desto leichter ist es natürlich, auf diese Entität zuzugreifen. Zudem kann man beide Dimensionen als Dimensionen der "Gegebenheit" bezeichnen: eine Entität, die für den Diskurs zentral ist, und auf die problemlos zugegriffen werden kann, ist sowohl für den Sprecher als auch für den Hörer "gegeben", und eine Entität, die erst noch in das Diskursmodell integriert werden muss, ist "neu".

E.3 Was ist die Struktur-Dimension?

Was die Struktur-Dimension des Entitätenstatus ist, das lässt sich nur im Rahmen einer geeigneten Theorie erklären, oder besser gesagt, explizieren. Um diese Explikation zu leiten, habe ich im ersten Abschnitt des Kapitels 3 drei Leitfragen definiert, die jede Theorie der Struktur-Dimension beantworten muss:

- In welchen Diskurssegmenten tritt die Entität auf?

  Dies setzt voraus, dass die Theorie so etwas wie Diskurssegmente definiert und einen Mechanismus bereitstellt, um das Auftreten von Entitäten in diesen Segmenten zu protokollieren. In der Diskurstheorie von Grosz und Sidner (1986) übernehmen die *focus spaces* diese Protokollfunktion. Ausserdem sollte die Theorie aufzeigen, in welcher Verbindung die Diskurssegmente zueinander stehen, in denen die Entität auftritt. In dieser Hinsicht bietet zweifelsohne die Rhetorical Structure Theory (Mann et al. 1992) die reichhaltigste Taxonomie an. Und was für Grosz und Sidner die *focus spaces* sind, das sind für die RST seit neuestem die Venen der Veins Theory.

- Wie ist die Entität mit anderen Entitäten im Diskurs verbunden?

  Dies ist der Aspekt, den alle betrachteten Theorien am wenigsten behandeln. Am ehesten kann man ihn mit Chafes (1994) Diskursthema oder dem Themenrahmen (*topic framework*) von Brown und Yule (1983) erfassen. Beide Konzepte schlagen Verbindungen von bereits erwähnten zu

potentiell erwähnbaren Entitäten aus demselben Themenkomplex.

- Wie eng ist die Entität mit den kommunikativen Absichten des Sprechers verbunden?

  Diese Frage kann nur durch eine adäquate Theorie der intentionalen Struktur von Texten geklärt werden. Die Theorie von Grosz und Sidner ist dazu unter den von mir untersuchten Theorien die beste Kandidatin, dicht gefolgt von der Arbeit von van Dijk (1980), der immerhin auch so etwas wie eine pragmatische Makrostruktur annimmt.

Abschnitt 3.4 hat klar gezeigt, dass wir nicht annehmen können, das Thema werde in jedem Diskurs irgendwann einmal direkt verbalisiert und somit explizit in den Status einer erwähnten Diskursentität erhoben. Dies gilt insbesondere dann nicht, wenn wir das Diskursthema so definieren wie Chafe oder Brown und Yule, nämlich als Komplex zusammenhängender Konzepte, Ereignisse und Zustände. Was wir sehr gut bestimmen können, ist jedoch, ob eine Diskursentität oft in bestimmten Segmenten vorkommt, oder wie oft sie die Rolle einer kontextuellen Verbindung gespielt hat. Das Auszählen von Vorkommen kommt dem quantitativen Begriff von Diskursthema recht nahe, den Levy (1982) entwickelt hat, und den Begriff “kontextuelle Verbindung” habe ich als Explikation des guten alten Satzthemas gewählt. Die Interpretation der quantitativen Ergebnisse ist einfach: je öfter eine Diskursentität in einem Abschnitt vorkommt, und je öfter sie als kontextuelle Verbindung fungiert, desto wichtiger ist sie für das Diskursmodell dieses Abschnitts.

Jedoch ist Thematizität nicht die Ebene, auf der ein Begriff des strukturellen Entitätenstatus definiert werden sollte. Obgleich der Begriff des Themas nützlich ist, um zu beschreiben, welche Rolle eine Entität in einem Diskurs spielt, so ist das doch nicht die Grundlegung, die wir eigentlich brauchen. Diese Grundlegung muss von einer Theorie der Diskursstruktur kommen. In Abschnitt 3.3 habe ich daraufhin drei Theorien miteinander verglichen, die Rhetorical Structure Theory (RST) von Mann und Thompson (1992), die Theorie von van Dijk (1980) und die Theorie von Grosz und Sidner (1986). Ein Sieger lässt sich nicht feststellen, obwohl die Strukturdimension des Entitätenstatus in allen drei Theorien expliziert werden kann. Die Theorie von Grosz und Sidner hat den Vorteil, dass sie gezielt die intentionale Struktur von Diskursen modelliert. Sie vernachlässigt jedoch darüber die semantische Makrostruktur, die van Dijk und die RST so detailliert erfassen können. Van Dijks Ansatz hat den Vorteil, dass er dank des Klassikers van Dijk und Kintsch (1983) eine Brücke zwischen psycholinguistischen Theorien des Textverstehens und der Textlinguistik schlägt. Mit van Dijks Superstrukturen lassen sich sehr schön textsortenspezifische Aspekte der Diskursstruktur beschreiben. Da van Dijk jedoch eine sehr detaillierte propositionale Darstellung von Texten als Eingabe verlangt, ist sein Ansatz für computerlinguistische Implementierungen mehr als ungeeeignet. Um jedoch psycholinguistische Experimente zu planen, deren Ergebnisse anhand von Walter Kintschs Konstruktions-/Integrationsmodell berechnet, simuliert oder interpretiert werden sollen, ist van Dijks Ansatz bestens geeignet.

Egal wie wir letztendlich die Struktur-Dimension explizieren, wir sollten sie nicht als Fundament der textuellen Kohärenz sehen. Zugegeben, referentielle Kontinuität ist wichtig; Diskursmodelle werden nun einmal um Diskursentitäten herum konstruiert. Jedoch gibt es genügend andere Ebenen, auf denen Kohärenz konstruiert werden kann. Solange dies dem Adressaten gelingt, sollten wir zufrieden sein.

E.4 Was ist die Verwaltungsdimension?

Die Verwaltungsdimension des Entitätenstatus kann man anhand von drei Aspekten beschreiben:

Initialisierung: Wie wird eine neue Diskursentität initialisiert? Wie ist sie mit dem Diskursmodell verbunden? Woher kommen wichtige Informationen, die dabei helfen, weiterhin auf die Diskursentität zugreifen zu können? Psycholinguistische Theorien des Verstehens sind sehr nützlich, um zu verstehen, was dabei vorgeht. Die Theorie mentaler Modelle (Johnson-Laird 1983) postuliert zum Beispiel, dass Menschen anhand analoger, unvollständiger, ständig sich ändernder mentaler Repräsentationen Perzepte verarbeiten, schlussfolgern, und Texte interpretieren. Dabei werden Perzepte aus verschiedenen Modalitäten integriert. Auf Elemente dieses Modells kann man sich rückbeziehen; diese Elemente fungieren dann als Diskursentitäten. Wie solche Elemente entstehen und in bestehende Modelle integriert werden, wie sie mit bestehenden Modellen weiter und enger verknüpft werden, das muss die allgemeine kognitive Theorie erklären, die hinter den Mentalen Modellen steht und von Johnson-Laird und seinen Schülern stets weiterentwickelt wird. Sanford und Garrod sind noch radikaler: Bei ihnen wird hereinkommendes Textmaterial sofort in ein Szenario integriert, ein sehr stark eingeschränktes Schema aus dem Weltwissen, ohne dass Sätze erst in Propositionen umgewandelt werden. Beiden Ansätzen ist gemeinsam, dass die Syntax nur als Hilfsmittel, als Quelle von Instruktionen zur

Integration in das mentale Modell / das Szenario dient, die wichtige Information mehr oder minder salient machen kann.

Zugriff: Beim Zugriff auf Diskursentitäten können verschiedene Informationsquellen angezapft werden: das Weltwissen, der Ko-Text, insbesondere mit der gerade spezifizierten Entität verbundene Diskursentitäten, Wissen über die Entität, das im Diskursmodell gespeichert ist. Ausserdem beeinflusst die *Salienz* einer Entität, wie schnell man auf sie zugreifen kann, und wie genau sie spezifiziert werden muss. Auf der letzteren Beobachtung hat Ariel (1990) eine weit greifende Theorie aufgebaut, die so genannte *Accessibility Theory*, mit der sie viele Phänomene der Verwendung referierender Ausdrücke erklären will; ja, *Accessibility* ist laut Ariel sogar die Dimension, entlang derer das System der referierenden Ausdrücke einer Sprache strukturiert ist. Die meisten Forscherinnen und Forscher teilen diesen Enthusiasmus nicht ganz, und unterscheiden Arten der Zugänglichkeit / Bekanntheit, wie z.B. Prince (1981). Chafe (1994) und Lambrecht (1994) unterscheiden zwei Hauptdimensionen: Identifizierbarkeit (ist die Diskursentität bereits im Modell oder aufgrund kon- oder kotextueller Information identifizierbar?) und Aktivierung (wie salient, wie zugänglich sind die identifizierbaren (potentiellen) Diskursentitäten? Chafe unterscheidet zwischen aktiver, semiaktiver und inaktiver Information. Das Problem mit dieser sehr kommunikativen Definition ist jedoch, dass sie sich nur schlecht an Korpora annotieren lässt, da gerade in der Grauzone der Übergänge zwischen den

Kategorien der Analyst sich genau in Sprecher und Hörer hineinversetzen muss, um zu ergründen, was hier möglicherweise so gerade noch für wen aktiv gewesen sein könnte. Eine explizitere und besser testbare, kognitiv sehr gut fundierte Theorie bietet dagegen Talmy Givón (1992,1995) an, dessen Überlegungen gut zu Kintschs (1988,1993) Modell des Textverstehens passen.

Update: Zum Schluss muss eine Theorie der Verwaltungsdimension erklären, wie die internen Repräsentationen von Diskursentitäten erneuert werden können, wie mit konfligierender Information umgegangen wird, wie mit Veränderungen einer Entität in der Zeit umgegangen wird, und über welche dieser Eigenschaften eine Entität noch erreichbar ist oder nicht.

E.5 Wie kann man Entitätenstatus in Korpora untersuchen?

Das fünfte Kapitel fasst einige grundlegende methodologische Überlegungen zusammen. Die Ausgangsfrage dabei ist: Wie kann man linguistische Korrelate des Entitätenstatus an Korpora untersuchen?

Um diese Frage zu klären, befasse ich mich zunächst mit Korpusstudien zum Thema Entitätenstatus. In Frage kommen Arbeiten, die korpusbasiert untersucht haben, wie Erst- gegenüber Zweiterwähnungen realisiert sind, korpusbasierte Arbeiten zum Thema Bridging, sowie korpusbasierte Arbeiten, die nachweisen wollen, dass man mit Begriffen wie “Bekanntheit”, “Zugänglichkeit” oder “Salienz” das Vorkommen bestimmter grammatikalischer Formen erklären kann. Einen Überblick über all diese Arbeiten zu geben, würde den Rahmen der Arbeit sprengen. Stattdessen konzentriere ich mich auf methodische Gemeinsamkeiten.

Eine Literaturanalyse ergibt, dass viele Arbeiten mit mehr oder minder beliebig herausgegriffenen Texten arbeiten. Nur selten werden eigens Korpora produziert, und wenn, dann sind das Korpora gesprochener Sprache. Viele Forscherinnen und Forscher setzen auf bereits annotierten Standardkorpora auf. Dies ist auch die durchgehende Strategie meiner Arbeit. Solch ein Vorgehen hat zwei Vorteile: Zum einen erspart es den Analysten wertvolle Zeit, die in eine tiefere Analyse der eigentlich zu betrachtenden Phänomene fliessen kann, zum anderen werden die Ergebnisse nachvollziehbar, da anderen Forschern dieselben Daten zur Verfügung stehen.

In dieser Arbeit stütze ich mich insbesondere auf zwei in der Computerlinguistik populäre Annotationsschemata: das MUCCS (Message Understanding Conference Coreference Scheme)-Schema (Girschman und Chinchor, 1997) und das MATE-Schema (Poesio 2000). Beide Schemata sind SGML- bzw. XML-basiert. Das MUCCS-Schema wurde für die Annotation von Nachrichtentexten entworfen. Es ist ausführlich validiert worden und bietet somit eine solide Basis für zuverlässige Annotationen. Daher bildete es die Grundlage für die Annotation aller Korpora in dieser Arbeit—allerdings mit einigen Änderungen. Die vielleicht wichtigste Änderung ist, dass entgegen dem ursprünglichen Schema Appositionen und prädikative Nominalphrasen nicht in Kospezifikationsfolgen eingefügt wurden. Das MATE-Schema wurde vor allem an die Annotation von Dialogen angepasst. Es stellt eine reichhaltige Taxonomie von Kategorien zur Annotation von Inferenzen zur Verfügung, die für die Interpretation mancher definiten NPs notwendig sind (engl.: *bridging*; Strube (1996) benutzt im Deutschen den Terminus “textuelle Ellipse”). Auf der Grundlage dieser Kategorien und der Arbeiten von Lambrecht (1994) und Prince (1981,1992) habe ich ein eigenes Annotationsschema entworfen. Dieses Schema kodiert, ob eine Diskursentität als “gegeben” angesehen werden kann, und wenn ja, aus welcher Quelle diese Informationen “gegeben” sind. Deshalb nenne ich das Schema “quellenbasiert”. Eine Übersicht findet sich in Tabelle E.1. Von dieser sehr ausführlichen Kodierung liessen sich bequem vier weitere, gröbere Taxonomien ableiten, diskursalt/-neu, höreralt/-neu, eine weitere,

	Quellenbasiertes Schema		**Abgeleitete Schemata**			
Kode	Kategorie	Beschreibung	STAT4	STAT3	DISC	HEAR
	brand new	dem Hörer unbekannt				
BU	*unanchored*	keine Verbindung zu existierender Diskursentität	BN	neu	neu	neu
BA	*anchored*	Verbindung zu existierender Entität	BN	med	neu	neu
U	**unused**	dem Hörer bekannt, dem Diskursmodell neu	U	alt	neu	alt
	accessible	erste Beschreibung kann konstruiert werden auf Basis von . . .	AC	med	neu	alt
SIT	*situation*	. . . Kommunikationssituation				
INF	*inference*	. . . Verbindung zu etablierter Diskursentität X durch—				
FRAME	frame:	Teil des durch X evozierten MOP				
PART	part/whole:	physischer Teil von X				
VAL	function/value:	Wert von X				
ISA	set (isa-Link):	Element von X				
SET	set (other):	Ober-/Untermenge				
EVENT	nominalization:	Nominalisierung der VP, die X denotiert				
AC	**active**	bereits erwähnt	A	alt	alt	alt

Tabelle E.1. Das quellenbasierte Annotationsschema und die davon abgeleiteten Taxonomien

hörerbasierte Taxonomie nach (Strube 1998), sowie eine grobe Unterteilung frei nach (Lambrecht 1994).

Solch eine detaillierte Taxonomie eignet sich natürlich nur für Texte, bei denen die Annotatoren eine einigermassen genaue Vorstellung vom Hörer haben; sonst ist es pure Spekulation, Quellen von Diskursentitäten annotieren zu wollen. Fehlt dieses Hörermodell, dann muss auf etwas zurückgegriffen werden, das auch an unbekannten Texten relativ gut und zuverlässig annotiert werden kann: Kospezifikationsfolgen, also Folgen von referierenden Ausdrücken, die dieselbe Diskursentität spezifizieren. Jeder dieser Ausdrücke “erwähnt” die

durch ihn spezifizierte Entität. Auf solchen Folgen lassen sich gut Distanzmaße definieren. Im letzten Abschnitt des fünften Kapitels untersuche ich verschiedene Aspekte solcher Maße. Als Erwähnung zähle ich bis auf weiteres für die Sprachen, die hier untersucht wurden (Englisch und Deutsch) nur explizite Erwähnungen als Nominalphrasen. Ich unterscheide anaphorische Grössen (wie z.B. Abstand zur letzten Erwähnung) von kataphorischen Grössen (wie z.B. der Persistenz). Ich argumentiere weiterhin, dass Distanzwerte, wenn überhaupt, theoriebasiert zu Variablen mit wenigen, sauber definierten Kategorien zusammengefasst werden sollten, da sich sonst unschöne Verzerrungen der gefundenen Werte und Probleme mit der Zuverlässigkeit der verwendeten Statistiken ergeben. Am ausführlichsten werden die *Einheiten* diskutiert, auf denen Abstandsmaße definiert werden müssen. Layoutbasierte Einheiten wie Absätze sind inhärent problematisch, da viele Faktoren beeinflussen, wann Autoren einen Absatz einfügen, nicht nur die thematische Kohärenz. Referierende Ausdrücke als Einheit sind ebenfalls problematisch, vor allem da sie im Fall komplexer Nominalphrasen recht komplexe Abstandsdefinitionen benötigen. Ein ähnliches Problem ergibt sich für Diskurssegmente, wo man bei Baumstrukturen mit Maßen aus der Graphentheorie wie der Länge des kürzesten Pfades zwischen zwei Knoten arbeiten kann. Die einfachste Lösung ist und bleibt, eine lineare Partition auf den Texten anhand syntaktischer Grenzen zu definieren. Diese Grenzen haben wir sehr weit gesetzt: Aus syntaktischen Gründen umfasst eine Einheit einen Hauptsatz, koordinierte subjektlose Hauptsätze, und

alle Nebensätze (Major Clause Unit, MCU, Strube und Wolters 2000).

Nachdem ich Kospezifikationsfolgen grundlegende Annotationskategorien definiert habe, versuche ich nun, diese Folgen zu modellieren. Dazu verwende ich zunächst eine recht grundlegende Art stochastischer Prozesse, die so genannten Poissonprozesse. Die Ergebnisse zeigen, dass solche Modelle zu starke Annahmen über die Verteilung der Erwähnungen einer Diskursentität machen. Die Abstände zwischen zwei Erwähnungen folgen nicht einer zeitinvarianten Verteilung. Ausserdem sind sie augenscheinlich nicht voneinander unabhängig—damit wäre eine wichtige Voraussetzung für viele Modelle stochastischer Prozesse nicht erfüllt. Letzten Endes wird man ein befriedigendes Modell nur erreichen können, indem man das Modell an eine stochastische Grammatik anbindet, und ein geeignetes Modell des Ko-Textes einbaut, vielleicht auf Grundlage von Kollokationen zwischen Diskursentitäten. Da das hier verwendete BROWN-COSPEC-Korpus viele verschiedene Themenbereiche abdeckt, lassen sich solche Kollokationen aus ihm nicht ableiten. Es bleibt weiterhin die Frage, ob sich Poissonprozesse tatsächlich zur Modellierung von Sequenzen von Erwähnungen eignen.

E.6 Empirische Exploration I: Radionachrichten

Im ersten Teil der empirischen Explorationen des Entitätenstatus, dem Kapitel 6, gehe ich zurück zu der Quelle meiner Überlegungen und Zweifel, zu der Textsorte der Radionachrichten. Mich interessiert zum einen, wie man bei dieser Textsorte überhaupt so etwas wie einen Entitätenstatus definieren kann, und zum anderen, was die linguistischen Korrelate dieses Status sind, so er sich als definierbar erweist.

Die Daten wurden nicht eigens für diese Arbeit gesammelt, sondern stammen aus Korpora, die in der Sprachforschung bekannt sind: dem Stuttgarter Radionachrichtenkorpus (Rapp 1998), und dem Bostoner Radionachrichtenkorpus (Ostendorf et al. 1995). Weiterhin wurden als Folie hinzugezogen das AUDIX-Korpus (Hirschberg 1993) prosodisch annotierter Agenturtexte sowie ein Korpus mit Radionachrichten eines öffentlichen und eines privaten Senders (Haaß 1994). Diese Vorgehensweise hat historische Gründe; ich bin durch meine Untersuchungen am Bostoner Korpus erst auf das Problem der Givenness aufmerksam geworden. Sie hat aber auch Vorteile für Prosodieforschung und Sprachtechnologieforschung. Da in beiden Bereichen viel mit meinen beiden Hauptkorpora gearbeitet wird, ist ein kritischer Blick auf die Textsorte sicherlich hilfreich, um die gewonnenen Ergebnisse besser beurteilen zu können.

Die medienwissenschaftliche Literatur zeigt deutlich, dass die Kommunikationssituation bei der Textsorte der Radionachrichten sehr verwickelt ist. Sowohl die Kommunikatoren als auch die Adressaten sind sehr heterogene Gruppen. Es ist also nahezu unmöglich, auf diesem Texten exakt zu definieren, was nun wirklich neue Information ist, und was man als bekannt voraussetzen kann. Als Kompromiss habe ich einen prototypischen Adressaten, Herrn Jupp Schmitz (englischer Vetter: John Doe), definiert, der sich durch politisches Desinteresse und leichte Simulierbarkeit auszeichnet. Aus der Sicht dieses Adressaten wurde der Status der Diskursentitäten bestimmt und annotiert.

Bei der eigentlichen Analyse der Daten beschränkte ich mich darauf, den Einfluss des Entitätenstatus auf die Wahl des Artikels und auf die Pronominalisierung zu untersuchen. Was Pronominalisierung angeht, so hat Entitätenstatus eine grosse Erklärungskraft. Die Variable, operationalisiert als Abstand zur letzten Erwähnung, kann mehr als 50% aller Variation in den Daten erklären. Wenn es jedoch um die Wahl des definiten Artikels geht, oder um die Frage, ob der Artikel weggelassen werden sollte, dann wird Entitätenstatus weit weniger wichtig. Als Haupt-Erklärungsprinzip dafür, wann welche Form von referierenden Ausdrücken benutzt wird, reicht er meines Erachtens nicht aus. Ich halte es auch für fragwürdig, den Entitätenstatus als funktionales Strukturprinzip für das System der referierenden Ausdrücke einer Sprache anzusetzen, wie Mira Ariel (1990) das gerne tut. Zwar beeinflussen kognitive Faktoren wie die von Chafe

herausgearbeiteten Faktoren der Identifizierbarkeit und der Aktivierung erheblich, wie eine Diskursentität spezifiziert wird; es gibt jedoch noch genügend andere semantische und soziale Faktoren, die dabei ebenfalls eine grosse Rolle spielen. Bei den Radionachrichten spielt sicherlich eine grosse Rolle, dass viel Information in wenig Zeit vermittelt werden soll. Das verführt dann dazu, in einen einzigen referierenden Ausdruck jede Menge elementarer Propositionen hineinzustopfen, die der Hörer erst wieder mühsam dekodieren muss, falls sie nicht bekanntes Wissen evozieren.

Bei einer detaillierteren Analyse stellte sich heraus, dass Pronomina im Vergleich zu Korpora wie BROWN-COSPEC eher selten vorkommen, und wenn, dann tritt ihr Antezedent meist im selben Satz auf. Definite Beschreibungen verhalten sich bemerkenswert neutral; ihr Vorkommen ist in einem solchen Maße unabhängig von ihrer Umgebung, dass man sie fast als unmarkierte Option für Radionachrichten bezeichnen könnte. Bei indefiniten Ausdrücken wiederum sieht die Lage anders aus. Sie sind eindeutig auf Ersterwähnungen spezialisiert, insbesondere, und dieses Ergebnis widerspricht einer oft in der Literatur aufgestellten Regel, wenn die erwähnte Entität nie wieder aufgegriffen wird.

NP level	
AGREE	Kongruenz (Numerus, Genus, Person)
Werte:	1 sg., 1 pl., 2 sg., 2 pl., 3 sg. masc., 3 sg. fem., 3 sg. neut., 3 pl.
SYN	Syntaktische Funktion
Werte:	Subjekt, Objekt, PP-Adjunct, andere
CLASS	Sorte
Werte:	siehe Tabelle C.2
Co-specification level	
SYNANTE	Syntaktische Funktion des Antezedenten
Werte:	erste Erwähnung in einer Kette, einzige Erwähnung, Subjekt, Objekt, PP-Adjunkt, andere
FORMANTE	Form des Antezedenten
Werte:	erste Erwähnung in einer Kette, einzige Erwähnung, Pronomen, Possessivpronomen, Demonstrativpronomen, definite NP, indefiniteNP (oder kein Artikel), Eigenname
DIST	Abstand zur letzten Erwähnung
Werte:	kein Antezedent im Diskurs, Antezedent in derselben MCU, Antezedent in der vorhergehenden MCU, Antezedent früher
PAR	Parallelismus
Werte:	kommt mit derselben syntaktischen Funktion im vorhergehenden Satz vor, ja/nein
COMPANTE	Ambiguität
Wert:	Anzahl kongruenter Diskursentitäten

Tabelle E.2. Übersicht über die verwendeten Faktoren. Alle Faktoren sind kategorial, außer COMPANTE, der ordinal ist

E.7 Empirische Exploration II: Pronominalisierung

In Kapitel 7 geht es um zwei Fragen: Wie gut kann struktureller Entitätenstatus, operationalisiert als Abstand zur letzten Erwähnung, erklären, wann eine Diskursentität mit einem Pronomen spezifiziert wird, und wann man lieber eine volle NP verwenden sollte? Kann man überhaupt genreunabhängige Pronominalisierungsregeln aufstellen, und wenn ja, welche Rolle spielt der strukturelle Entitätenstatus dabei? Wir haben uns dabei auf den grundlegendsten Fall beschränkt: wir wollten alle Pronomina vorhersagen, egal welcher Person. Dies

ist ungewöhnlich, weil sich viele vorherige Studien auf Personalpronomina der dritten Person beschränkt haben; manche Studien haben auch noch die Diskursentitäten auf solche eingeschränkt, deren Referent ein Mensch ist.

Beide Fragen wurden auf dem BROWN-COSPEC-Korpus untersucht. Dabei wurden zwei Arten von Verfahren eingesetzt. Mit Hilfe der logistischen Regression können wir detaillierte Hypothesen darüber testen, welche Faktoren die Pronominalisierung beeinflussen, und wir können ermitteln, wie stark dieser Einfluss ist, und wie stark die Faktoren miteinander wechselwirken. Durch den Einsatz von Algorithmen zum maschinellen Lernen können wir klären, inwiefern die Faktoren ausreichen, um vorherzusagen, ob eine Erwähnung als Pronomen realisiert wird oder nicht. Wir verwenden dazu einen regelbasierten Lerner, RIPPER (Cohen 1995), und einen instanzenbasierten Lerner, IB1-IG (Daelemans, Zavřel, van der Sloot and van den Bosch 1999). Streng genommen ist auch die logistische Regression ein Vorhersageinstrument, so dass wir am Ende drei Ansätze zum maschinellen Lernen von Pronominalisierung vergleichen können: statistische Schätzverfahren (logistische Regression), Regelinduktion (RIPPER) und beispielbasiertes Lernen (IB1-IG).

Die erste der beiden Fragen kann bejaht werden: Das Abstandsmaß, das hier definiert wurde, kann tatsächlich viele Vorkommen von Pronomina erklären. Wir untersuchten, ob weitere Faktoren die Leistung verbessern. Diese Faktoren sind in Tabelle E.2 zusammengefasst. Wir beschränkten uns dabei auf solche

Faktoren, die entweder einfach aus vorhandenen Annotationen extrahiert werden können, oder die zuverlässig annotiert werden können. Unsere Motivation dafür war nicht, einen "wissensarmen" Ansatz zur Generierung von Pronomina zu entwickeln. Unsere Motivation war eher theoretischer Natur: statistische Tests auf unzuverlässig annotierten Daten sind wenig sinnvoll.

Zur zweiten Frage konnten wir ebenfalls eindeutige Ergebnisse erzielen. Welcher der Faktoren wie wichtig für die Pronominalisierung ist, das hängt eindeutig vom Genre ab. Der einzige wirklich robuste (aber wenig erklärungsstarke) Faktor war COMPANTE, Ambiguität. Wir können zwar eine Menge von vier Faktoren ausmachen, die für alle Genres passable Ergebnisse liefern: DIST, COMPANTE, FORMANTE und AGREE. Diese Faktoren liefern jedoch nicht immer die besten Ergebnisse.

Die detaillierte Sorten-Ontologie, die in Anhang C.2 beschrieben ist, war leider wenig hilfreich. Die wichtigsten Unterscheidungen waren, wie aus der Literatur zu erwarten, [$\pm$ belebt] und [$\pm$ abstrakt]. Die syntaktischen Faktoren PAR, SYN, und SYNANTE waren bei weitem nicht so wichtig wie die Literatur suggeriert. Wir denken nicht, dass dies daran lag, dass unsere Einteilung in syntaktische Funktionen zu wenig detailliert war; sie kommt der ebenfalls sehr groben Einteilung von Givón (1992) recht nahe.

Curriculum Vitae

Name	Maria Klara Wolters
Eltern	Heinz-Dieter Wolters, geb. am 22. 4. 1943
	Roswitha Wolters, geborene Gingter, geb. am 12. 8. 1943
geboren am	26. 3. 1974
Geburtsort	Rheydt
Familienstand	ledig
1980–1984	Katholische Grundschule Konstantinstraße, Rheydt-Giesenkirchen
1984–1992	Franz-Meyers-Gymnasium, Rheydt-Giesenkirchen
Juni 1992	Abitur
1. 10. 1992–	Studium der Kommunikationsforschung und Phonetik, Informatik und Allgemeinen Sprachwissenschaft an der Rheinischen Friedrich-Wilhelms-Universität Bonn
1. 10. 1993–	Studium der Informatik an der Rheinischen Friedrich-Wilhelms-Universität Bonn
25. 11. 1997	Abschluß: Diplom-Informatikerin
	Thema der Abschlußarbeit: A Diphone-Based Speech Synthesis System for Scottish Gaelic
1995/96	Studienaufenthalt an der University of Edinburgh, Department of Linguistics, Centre of Cognitive Science, Department of Celtic Studies
1993–1998	Stipendiatin der Studienstiftung des Deutschen Volkes
1994–1995	studentische Hilfskraft am
und 1996–1997	Institut für Kommunikationsforschung und Phonetik der Rheinischen Friedrich-Wilhelms-Universität Bonn
Dezember 1997–	wissenschaftliche Mitarbeiterin am Institut für Kommunikationsforschung und Phonetik